RICE UNIVERSITY

Network-guided genome-wide studies reveal a complex genetic architecture of warfarin resistance in the Norway rat (*Rattus norvegicus*)

by

Shuwei Li

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Michael H. Kohn, Chair
Associate Professor of Ecology and Evolutionary Biology

Nicholas Putnam, Assistant Professor of Ecology and Evolutionary Biology

Luay K. Nakhleh, Associate Professor of Computer Science

Marek Kimmel, Professor of Statistics

Bo Peng, Assistant Professor of Epidemiology

HOUSTON, TEXAS
February 2013

ABSTRACT

Network-guided genome-wide studies reveal a complex genetic architecture of warfarin resistance in the Norway rat (*Rattus norvegicus*)

by

Shuwei Li

A fundamental challenge in evolutionary biology and medical genetic research is to connect the phenotype (a disease in humans or an adaptive trait in animals or plants) with the genotype. Using a classical example of an adaptive trait with a strong Mendelian genetic basis - warfarin resistance in the Norway rat (*Rattus norvegicus*), my dissertation tests the main hypothesis that speculated 'simple' adaptive trait has a more complex genetic architecture.

Warfarin is an anticoagulant rodenticide used since the 1950s, and also is a widely prescribed blood-thinning drug in human. As a rodenticide, warfarin has initially been very effective. However, resistant rodents have evolved quickly and *Vkorc1* (vitamin K epoxide reductase complex subunit 1) is the known resistance gene. As a popular drug, warfarin has a narrow therapeutic window with several genes *VKORC1*, *CYP2C9*, *CYP4F2* established as biomarkers predicting warfarin dose in humans, suggesting a complex genetic architecture of warfarin resistance in rodents.

In my thesis I performed network-guided genomic association studies (NetGWAS) and gene expression analysis to identify candidate genes involved in warfarin resistance based on a sample of ~600 wild rats from 19 populations in Germany.

My thesis work revealed that the resistance mutation in *Vkorc1* likely is under balancing selection and was recently introduced to the rat population in our study area.

A key innovation of my thesis is adopting a NetGWAS approach to prioritize true associations and conducting co-expression network analysis to detect expression changes related to warfarin. My work shows that additional candidate genes are connected to the vitamin K pathway of which *Vkorc1* is an essential component. While the validation of identified genes remains a challenge, the value of my thesis for future investigation is shown: one candidate gene *Calu* (Calumenin) is associated with warfarin resistance in multiple populations and is an essential part of the vitamin K cycle. Finally, my thesis briefly examines the genetics underlying a newly postulated cost of resistance, arterial calcification.

This dissertation provides us an innovative framework in which we learned the genetic architecture of an adaptive trait in multiple dimensions: nucleotide or expression variation, genomic distribution and gene-gene interactions.

Acknowlegements

Aug 04, 2007, first time stepping on US, Houston, Rice, I was exciting and nervous. Even it was a Saturday, Michael immediately drove from home to meet me after receiving my call; even I didn't fully understand the German-style English and the jokes, I immediately knew he would take care of me and be a good mentor not only for my PhD. 6 years later, I feel really lucky for knowing and working with so many insightful and incredible people. The fortune started with my advisor, Michael Kohn, who has done every possible thing to teach, support and guide me through the whole process with great patience. Especially, Michael, I thank you for encouraging me to develop research ideas and helping me grow as a scientist, which will benefit me for my whole life.

Many thanks to my committee: Dave Queller, Nicholas Putnam, Luay Nakhleh, Marek Kimmel and Bo Peng. Each year, I learned invaluable advices and helpful critiques from you; each year, I feel stupid for not being able to answer questions and then motivated to do a better job next time. Moreover, I learned and grew from the courses you taught, the codes you edited and the opportunities and the help you provided for me on my research and career.

Many thanks to all my colleagues: Chinghua Shih, Ying Song, Juan Diaz, Carlos Nossa, Xiaoyun Liao. I greatly appreciated knowing, working with and discussing serious research or having fun with you. I enjoyed our big laughs, the constructive arguments, the time spending together on pipetting, coding, reading, encouraging each other and sharing.

I learned a lot from each of you and can never be thankful enough for all of your supports. My special thanks go to my husband and colleague, Zhenjiang Lan, who is on call 24/7 to solve problems with me or for me, from wet lab to dry lab, from hardware to software, from building database/website, processing raw data, algorithm implementation to result visualization and presentation, including drawing cartoons. A lot of times I felt so mean to ask for result in 5 minutes or in midnight, and you did it; a lot of times, I proposed a 'crazy' idea or expectation and you realized it. This cannot be accomplished without you.

I would also like to express my gratitude to people at our department, especially to Laura Johnson, Diane Hatton, Lesley Campbell, Evan Siemann, Zhenguo Lin, Jiaxing Yue, Juli Carillo, Amy Dunham, Tom Miller and Rachel Stones for always being caring and supportive. I gratefully acknowledge the financial and computing supports from Michael, EEB department, Rice University and NIH.

My sincere thanks go to my family and friends for accompanying me and helping me overcome difficulties in life. Especially thank Lily Lam from OISS and all RCCFs, I may not be here today without you. I know it is not just luck, I thank God, mom, dad, Lan and my little baby Timo; this is dedicated to you, my loved ones.

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

BCR	Blood Clotting Response
CTD	Comparative Toxicogenomic Database
GGI	Gene-gene Interaction
GO	Gene Ontology
GWAS	Genome wide association study
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
QTL	Quantitative Trait Locus
eQTL	Expression quantitative trait locus
SNP	Single Nucleotide Polymorphism

List of Gene Names

ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1
APOE	Apolipoprotein E
Bglap	Bone gamma-carboxyglutamate (gla) protein
Bmp2	Bone morphogenetic protein 2
Calu	Calumenin gene
CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit
Cacna2d1	Calcium channel, voltage-dependent, alpha2/delta subunit 1
Cacna2d3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3
CYP1A1	Cytochrome P450, family 1, subfamily A, polypeptide 1
CYP1A2	Cytochrome P450, family 1, subfamily A, polypeptide 2
CYP2C8	Cytochrome P450, family 2, subfamily C, polypeptide 8
CYP2C9	Cytochrome P450, family 2, subfamily C, polypeptide 9
Cyp2c11	Cytochrome P450, subfamily 2, polypeptide 11
CYP3A4	Cytochrome P450, family 3, subfamily A, polypeptide 4
CYP3A5	Cytochrome P450, family 3, subfamily A, polypeptide 5
CYP4F2	Cytochrome P450, family 4, subfamily F, polypeptide 2
Ephx1	Epoxide hydrolase 1, microsomal gene
F2	Coagulation factor II
F7	Coagulation factor VII
F9	Coagulation factor IX

F10	Coagulation factor X
Fgfr2	Fibroblast growth factor receptor 2
FGFBP2	Fibroblast growth factor binding protein 2
GAS6	Growth arrest-specific 6
Ggcx	Gamma-glutamyl carboxylase gene
MGP	Matrix Gla protein
Nqo1	NAD(P)H dehydrogenase, quinone 1 gene
NR1I2	Nuclear receptor subfamily 1, group I, member 2
NR1I3	Nuclear receptor subfamily 1, group I, member 3
ORM1, ORM2	Orosomucoid 1, Orosomucoid 2
PROC	Protein C, inactivator of coagulation factors Va and VIIIa
PROS1	Protein S (alpha)
PROZ	Protein Z, vitamin K-dependent plasma glycoprotein
SERPINC1	Serpin peptidase inhibitor, clade C (antithrombin), member 1
Vkorc1	Vitamin K epoxide reductase complex subunit 1

Chapter 1

Introduction

What is life? In the last paragraph of the book *On the Origin of Species by Means of Natural Selection*, Darwin said: “life is which evolves” (Darwin 1859). 150 years later, J. Craig Venter described it in another way: “life is a software system and DNA is the genetic code”. As suggested by Jerry A. Coyne and others, what Darwin really told us in his book is not the origin of species but the origin of adaptation (Coyne 2009). Now with better knowledge and more advanced tools in hand, Dr. Ventor’s definition urges us to understand what exactly happened during the adaptation process, i.e. what is the precise nature and role of changes in DNA that affect evolving systems. Since Darwin emphasized that “many slight differences” are the source of variation for selection to act upon, generations of scientists have been examining the question of “what is the genetic basis of adaptation” (Orr 2005; Hurst 2009; Wang et al. 2011; Radwan and Babik 2012; Axelsson et al. 2013; Linnen et al. 2013).

Interestingly, this search for general answers that explain adaptive trait evolution in natural populations of animals and plants parallels to the search for answers regarding the genetics underlying disease in humans. The enduring effort in medical genetic research to identify genes and pathways, and genetic mechanisms that explain disease has been a powerful driver of molecular tools and theoretical innovation, and recently, genomic tools and genome-wide association analyses (Lucia A. et al. ; Risch and Merikangas 1996; Consortium 2005; Schadt et al. 2005; Consortium 2007; Purcell et al. 2007; Gilad, Rifkin, and Pritchard 2008; Slatkin 2008; Takeuchi et al. 2009; Barabasi, Gulbahce, and Loscalzo 2011; Liu et al. 2011; Stranger, Stahl, and Raj 2011; Li et al. 2012).

1.1. Background

"We know little about the genetic basis of adaptation" (Orr and Irving 1997). Is the adaptation built from a few genes of large effect or many genes each of small effects? Does adaptation occur from *de novo* mutation or standing variation? What is the mode of selection on the mutant alleles and are these recessive, dominant, or over-dominant? New technologies have been developed and considerable progress has been made with regard to technological and theoretical developments; but we still know little about the genetic basis of adaptive traits in natural populations of animals and plants, and with the advent of genome-wide perspectives and data we are faced with as many new opportunities as we are faced with new challenges. For instance, we now are able to examine how does selection shape the genome as a whole? If

multiple selective events are discovered we might wonder whether these are part of a single selective event affecting multiple loci of if each selected site is under its own selective pressure? In this context selection is broadly defined, but potentially might need to be viewed in a refined fashion, such as selection on driver mutations initially leading to an adaptation which are followed by mutations enhancing the trait and mutations compensating for fitness costs associated with driver mutations. From a genome-wide perspective we can pose the general question whether adaptation occurs at the levels of amino acid sequences or at the level of gene expression? A key question underlying my thesis is whether the potentially complex genetic architecture of adaptive traits as such is simpler once we adopt a gene-gene interaction network perspective. In other words, perceived complexity in terms of the number of genes involved might be reduced to single of few pathways. Within such pathways a challenge would be to distinguish genes that directly interact with selective agents from genes that are compensating for any fitness costs.

Understanding the genetic basis of adaptation requires us to bridge the gap between genotype and phenotype. As defined by Wilhelm Johannsen in 1911, the "genotype is an organism's heredity and phenotype is what that heredity produces." (Johannsen 1911). Bridging the gap between genotype and phenotype requires the identification of genetic variations that contribute to the phenotypic trait. Moreover, it requires testing for the strength, mode and timing of selection for each variant. A novel research direction in this context involves the exploration of any gene-gene interactions and functional relationships among variants. In Figure 1.1 I summarize some of the steps involved in contemporary and future studies with aim to bridge the

genotype and phenotype gap. First, genes and other variants responsible associated with a phenotype are isolated by genome wide association analyses of based on other mapping strategies. The trait or disease of interest might be determined by one primary gene or be complex in its architecture. Note, however, that based on my thesis work I propose that adaptation in nature may not seem to progress by simple genetics; instead, simple genetics progress towards complex genetics over time through recruitment of new mutations and standing variants to either enhance the trait, buffer the trait, or compensate for a cost of the main driving mutations.

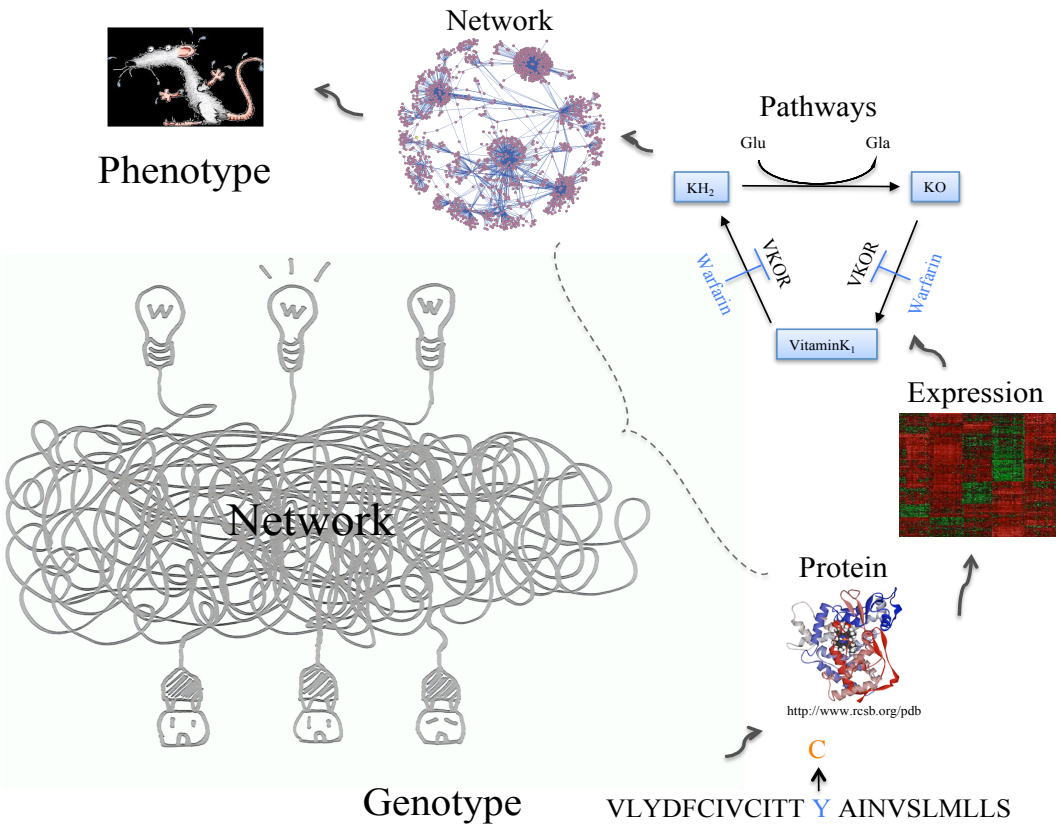


Figure 1.1 – Linking the genotype with the phenotype by examining nucleotide variation and experssion changes with a network perspective as genes or proteins are interacted with each other in pathways and networks.

Genome wide variation and association studies (GWAS), as well as analyses of gene expression have revealed a wealth and diversity of genomic variation, and clearly have shown that mutations on DNA sequences frequently result in altered protein structures and gene expression patterns to the effect that these affect function (Figure 1.1). Moreover, systems biological perspectives and tools applied to genomic data clearly revealed that genes are not working in isolation, but the genome should be viewed as a network connecting subnetworks (pathways), and connections (edges) in the network can be encoded in numerous ways, including as gene expression changes or protein-protein binding, just to mention a few. Thus, if we zoom out from the resolution at single genes, individual pathways, to the thousands of interacting genes, it becomes clear that for function to be maintained individual components need to interact properly, or are well buffered to change. Moreover, the disruption caused by new mutations, often of dramatic effect such as the Y139C mutation (Y to C amino-acid change at position 139 of the protein) in *Vkorc1* that I study, can be anticipated by examining genetic networks (Rost et al. 2004). In addition, this network perspective raises questions regarding a simplistic view of single selective events under study in natural populations. For instance, even an initially single gene response could result in cascade effects in the network, and thus, result in what appears as multiple selective events. If these selective events are detected and interpreted in isolation the importance of selection on the genome would appear inflated, as single events result in numerous responses at the genome level. For these, and many other reasons, adopting a network perspective is needed to more effectively explore the fundamental questions concerning the genetic basis for adaptation.

In my thesis I study a presumably very simple model system: warfarin resistance in wild rats (*Rattus norvegicus*). The rationale to study such a presumably simple and unusual pesticide resistance system is to illustrate that even in such system complexity is found. This should provide guidance to studies of natural study systems that obviously are more complex. Results of my work suggest that the current state-of-the art of interpreting genomic data collected on wild populations appears to result in misleading conclusions in that a mere tabulation of genomic regions under selection only is a small part of studies that aim to understand the genetics of adaptation.

Warfarin (a derivative of 4-hydroxycoumarin) was introduced in the 1950s as an anticoagulant rodenticide. It works by blocking the vitamin K 2,3-epoxide reductase complex (VKOR) of the vitamin K cycle, and subsequently impairing the vitamin K-dependent gamma-carboxylation system and the following activation of blood coagulation factors, and, as a presumably unnecessary side effect, should reduce the activities of vitamin K-dependent proteins not involved in blood coagulation (c.f. Appendix 8, Figure 5.2) (Presnell and Stafford 2002; Stafford 2005). Thus, susceptible rodents succumb to internal hemorrhage after ingestion of the poison and the 'blood-thinning' effect of warfarin made the substance extremely useful as a rodenticide. In addition, it was quickly realized that this blood thinning effect could be taken advantage of by turning warfarin into a drug to prevent heart attack, stroke or thrombosis in human (Gage et al. 2003; Geisen et al. 2005).

While my thesis is mainly concerned with the evolutionary genetics of adaptive evolution of a trait, my thesis has important implications to the biomedical sciences also. Specifically, although warfarin has been widely prescribed as a drug for about 60 years, it remains a difficult drug to manage because of its narrow therapeutic range, wide dosage variation, and serious side effects if dosed inappropriately. The optimal dosage varies dramatically within and between human patients owing to genetics, diet, drug-drug interactions, and other environmental factors (Kamali 2006). It is thus a challenge to determine a safe yet effective dose of warfarin for individual patients that prevents thrombosis while avoiding bleeding risks (Takeuchi et al. 2009). In humans, several mutations in the genes *VKORC1* (encodes VKOR), *CYP2C9* (cytochrome P450, family 2, subfamily C, polypeptide 9; encodes enzyme metabolize the S-enantiomer of warfarin), and *CYP4F2* (cytochrome P450, family 4, subfamily F, polypeptide 2; vitamin K_1 oxidase) affect the physiological response to warfarin drug treatment, and thus, presently these are established as the only biomarkers predicting warfarin dosage (Li et al. 2004; Rost et al. 2004; McDonald et al. 2009; Pautas et al. 2009; Takeuchi et al. 2009). Other candidate genes such as *GGCX* (Gamma-glutamyl carboxylase) and *EPHX1* (Epoxide hydrolase 1, microsomal) were reported, but with controversial results obtained during different studies (Wadelius et al. 2005; Cha et al. 2007; Wadelius et al. 2007; Pautas et al. 2009). Nonetheless, incorporating genetic information greatly improved the warfarin dosage estimation, therby averting an estimated 85,000 serious bleeding cases and 17,000 strokes annually. Genetically informed warfarin dosing has dramatically reduced health care costs (~ 1 billion annually, source: AEI-Brookings joint center for

regulatory studies). Now genetic forecasting of warfarin dose is actively studied and practiced (IWPC ; Wadelius et al. 2009; Pavani et al. 2012; UI-Health 2012). Even though warfarin has become a poster child for the future of personalized medicine my thesis shows that warfarin dosing in this application area would benefit from an even more comprehensive understanding of the genetic underlying physiological responses to warfarin. Such searches for additional genes are complicated in humans and potentially could be more effectively done in our study system of free-ranging rats that have been inadvertently selected for decades to resist the poison. In this system we would expect that any variant and gene that enhances the fitness of rats theoretically had a chance of fixation of $2N_es$, as opposed to $1/N_e$ in human populations where warfarin response is not a selected trait. Interesting, my thesis results indicate that drift plays a larger role in the study system that was anticipated at the onset of the study, however.

As the first generation rodenticide, warfarin was initially a highly effective tool to control rats (Endler 1985; Gillespie 1991; Kohn, Pelz, and Wayne 2000). However, resistance to warfarin has evolved within a mere ~10-15 years after its introduction in the 1950s. Warfarin resistant rats have been found, first in Scotland (1958) and subsequently at the Wales-Anglo border (1960), Denmark (1962), Holland (1966) and Germany (1967) (Boyle 1960; Lund 1964; Jackson and Kaukeinen 1972; population and agriculture 1986). Now resistant rats have been found in many locations dispersed across the globe, with much of resistance yet to be discovered because systematic surveys rarely being conducted (Figure 1.2).

Figure 1.2 – Distribution of warfarin resistance for Norway rats (*R. norvegicus*) around the world. Map is generated in GIS (Geographic Information Systems).

Resistance to warfarin was speculated to have simple genetic basis (Greaves and Ayres 1969; Wallace and MacSwiney 1979). It is currently accepted that *Vkorc1* is the major gene that causes warfarin resistance in rodents (Kohn and Pelz 1999; Kohn and Pelz 2000; Li et al. 2004; Rost et al. 2004; Pelz et al. 2005). The best-documented case where resistance by *Vkorc1* is conferred is by a Y->C substitution at position 139 of VKORC1 that decreases warfarin binding affinity (Kohn, Pelz, and Wayne 2003; Rost et al. 2004; Rost et al. 2009). This Y139C mutation occurs in our study area in Germany. The vertebrate homologues to *Vkorc1* gene are found in arthropods, plants, bacteria and archaea (Goodstadt and Ponting 2004), suggesting the gene is involved in pathways and physiological/developmental processes other that blood coagulation.

1.2. Rationale

Norway rats in our study area in Germany have experienced decades of strong selection with the rodenticide warfarin. While it is commonly thought that at least three genes (*VKORC1*, *CYP2C9* and *CYP4F2*) are associated with warfarin dosage variance in human, and *Vkorc1* in rats, a main question of my thesis is whether the strong and persistent selection with warfarin has selected for additional genes in the rat genome (Figure 1.3). If this main working assumption of the thesis is met then the fundamental question studied is whether these genes are clustered in terms of function and pathways in a gene-gene interaction network. If this hypothesis was true then the main contribution of the thesis, in terms of evolutionary theory, would be i) that even simple genetic adaptations mentioned in evolutionary textbooks and literature might in fact more commonly be more complex. ii) Complexity of adaptive traits should be interpreted not with regard to the number of genes involved alone, but also within the context of the number of pathways involved.

Moreover, as I discuss in detail in Chapter 2, the Y139C resistance mutation imposes fitness costs (reduced growth rate and arterial calcification, etc.) to rats. Thus it is interesting to explore the existence of other genes with potentially compensatory mutations that have become a part of the genetics of warfarin resistance.

Figure 1.3 – Rationale of thesis. Several genes *VKORC1*, *CYP2C9*, *CYP4F2* are established as biomarkers predicting warfarin dose in humans. We expect to detect additional genes other than *Vkorc1* related to warfarin resistance in rodents which have experienced strong selection.

1.3. Overview of methods

In this dissertation project, I performed a network-guided genomic study utilizing innovative computational and statistical tools to search for genetic variants associated with warfarin resistance, and to investigate these variants in terms of selection and in terms of gene-gene interactions that they might mediate or affect.

The first focus is on the known resistance gene *Vkorc1*. While it is commonly assumed that the warfarin resistance trait is under balancing selection as mediated by overdominance, this has not been confirmed by any population genetic analysis of the resistance gene *Vkorc1*. Interestingly, patterns of linked microsatellite variation indicated directional selection rather than balancing selection at the locus (Kohn, Pelz, and Wayne 2000). Thus, the first sets of methods used include SNP typing at *Vkorc1* and linked sites and population genetic analyses. Forward time simulations are used to distinguish competing hypotheses regarding timing, strength, and mode of selection at *Vkorc1* consistent with the observed patterns.

In the following I switch from a gene-centric perspective to a genome-wide perspective. Notably, in our laboratory the candidate gene-centric perspective has been applied to the two other obvious candidate genes of the cytochrome 450 types. To learn the overall genetic architecture underlying warfarin resistance the rat genome is searched for SNPs associated with resistance. This is done on an Affymetrix SNP typing platform in collaboration with genome centers. Identified variants and the haplotypes they are located on are examined selection and undergo functional annotations. Designed to identify genetic variations associated with complex diseases and traits, genome wide association studies (GWAS) have gained tremendous attention (Lucia A. et al.). I detected significant SNPs by measuring association strength as Bayes Factors during Bayesian association analyses. However, this traditional GWAS approach detected candidate variants individually and independently, resulting in large lists of candidate genes, many of which likely to be false positive owing to linkage to functional sites, and GWAS did not account for

interesting gene-gene interactions as part of the significance assessments (Cordell 2009). Hence we expanded the traditional GWAS to Network-guided GWAS (NetGWAS) by evaluating the significance of association of genes with the trait in the context of their location in gene-gene interaction networks. Thus, trait relevant genes are prioritized. A key innovation of this thesis is the implementation of a modified Google's PageRank algorithm to conduct these analyses (Page et al. 1999). The larger set of candidate genes is analyzed with population genomic tools, and forward time simulations are used to test null models predicting the randomly expected number of genes underlying the trait as explained by drift (and chance) alone. The thesis addresses another novel challenge neglected in data-driven molecular population genetics analyses, which is to detect selection on sets of interacting genes and within the frameworks of genetic network.

The ability to conduct analyses of gene expression added an important "intermediate level" that might help filling the genotype and phenotype gap, and on occasion is referred to as the closest proxy for a phenotype that is a target of selection (Ellegren et al. 2012). Some variations, like the resistance mutation on *Vkorc1* gene, change the protein structure (Kohn, Pelz, and Wayne 2003; Rost et al. 2009). Some variants, on the other hand, might alter the gene expression, such as the mutations in the promoter region of the lactase gene underlying the adaptation of humans to milk consumption (Tishkoff et al. 2007). Previous studies have reported few genes with differential expression patterns on small custom arrays that are associated with rodenticide resistance (Markussen et al. 2008a; Markussen et al. 2008b). Here the Affymetrix platform was used to obtain genome-wide expression profiles for wild

rats kept in the laboratory of collaborators testing these for resistance status. Such testing involves the injection of rats with low doses of warfarin. Thus, we assay gene expression for resistant and susceptible rats induced and non-induced with warfarin. These data are used to search for associations between gene expression and the resistance trait, as well as to construct co-expression networks. In this thesis the SNP association results and gene expression results are combined for a more comprehensive view of the genetics underlying warfarin resistance, and while considering candidate genes. Methods employed include an analysis of population structure to account for stratification. Finally, I collected further genotype data for larger population samples for top candidate genes to better understand the population genetics of these genes, including the effects of selection and drift.

1.4. Outline of the Dissertation

After an introductory Chapter 1 thesis Chapter 2 focuses on a population genetic analysis of the *Vkorc1* gene and linked polymrophisms in a wild-derived population NW and a non-resistant population LH. We connected genotype, phenotype and fitness by analyzing the association of the Y139C underlying SNP and evaluated various models of selection n the gene and haplotype it is located on. Three fundamental questions are examined in Chapter 2. The strength and mode of selection at *Vkorc1* are analyzed and the age of the mutation is estimated.

Chapter 3 is concerned with a network-guided GWAS (NetGWAS) on a 10k rat SNP array on which a wild-derived population, NW, and a non-resistant

population, LH, were assayed with aim to detect associations of SNPs with warfarin resistance (SNP array I). I combined the gene-gene interaction (GGI) network information with the gene-trait association measures based on a modified Google's PageRank algorithm (Page et al. 1999; Morrison et al. 2005). Gene ranks are given to prioritize warfarin related genes followed by functional annotation analyses based on the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information (Ogata et al. 1999; Consortium 2008). Population genomic analyses were performed between the resistant and non-resistant populations, including tests for extended haplotypes and linkage disequilibrium (LD).

Chapter 4 describes the work involved to find genes that display association of their expression with the resistance phenotype under various experimental variables (induction with warfarin, gender, and warfarin resistance genotype at *Vkorc1*). Co-expression networks are built based on the microarray data collected from the same wild-derived populations, NW and LH, which were used in Chapter 3. A stunning 600 genes that were correlated with warfarin treatment and resistance phenotype emerged, which clearly was unexpected given that warfarin was supposed to be a pathway-specific drug and poison. This observation was critically evaluated and in chapter 3 it is shown that linkage of many polymorphic variants underlying expression polymorphisms in wild rats must be linked to the causal sites under selection, and genetic hitchhiking and neutral haplotype structure cased these to magnify the gene expression response to warfarin. Network analyses in functional terms were crucial steps in the identification of candidate genes, identifying 21 clusters. Although the known resistance non-synonymous mutation underlying

Y139C is not known to modify the expression of *Vkorc1* gene, the surrounding genes exhibited signals of warfarin related expression variations and formed a detectable cluster. This finding showed the promises of other clusters to form false-positive associations by random chance (linkage). Results from chapter 2 were used to refine the analyses of gene expression by mapping SNP array data onto the annotated gene expression data. This reduced the number of subnetworks to 7. Finally, by eQTL mapping (expression quantitative trait loci) we increased our knowledge about the architecture of gene regulation (Gilad, Rifkin, and Pritchard 2008).

In Chapter 5, we performed the network-guided genomic association study (NetGWAS) of the rat 10k SNP array data in multiple additional wild populations (SNP array II). We used population structure analysis of ~ 600 rats to select 46 samples suitable for the SNP array experiment (SNP array II) in that they are part of the same (non-stratified) clusters identified. For discovery and validation purposes, genes with multiple supporting evidences from all chapters were reported at the top of this candidate gene list. For one candidate gene, *Calu*, the association was validated in a sample of ~600 rats.

In Chapter 6 some key implications of this thesis are discussed and future work on this specific study system are provided, as well as some general recommendations for the genomic study of the genetic architecture underlying adaptive traits (Figure 1.4).

1.5. Contribution of the Dissertation

This dissertation built a framework for understanding the genetic basis of warfarin resistance (Figure 1.4), which utilized genomic platforms, bioinformatics, and population genetics to test the main hypothesis that even adaptive traits once thought to be simple in their genetics have a more complex genetic architecture. In this framework, we first tested selection at the genetic level for the known resistance gene *Vkorc1*, which would serve as an example for decoding signals of selective sweep (Chapter 2).

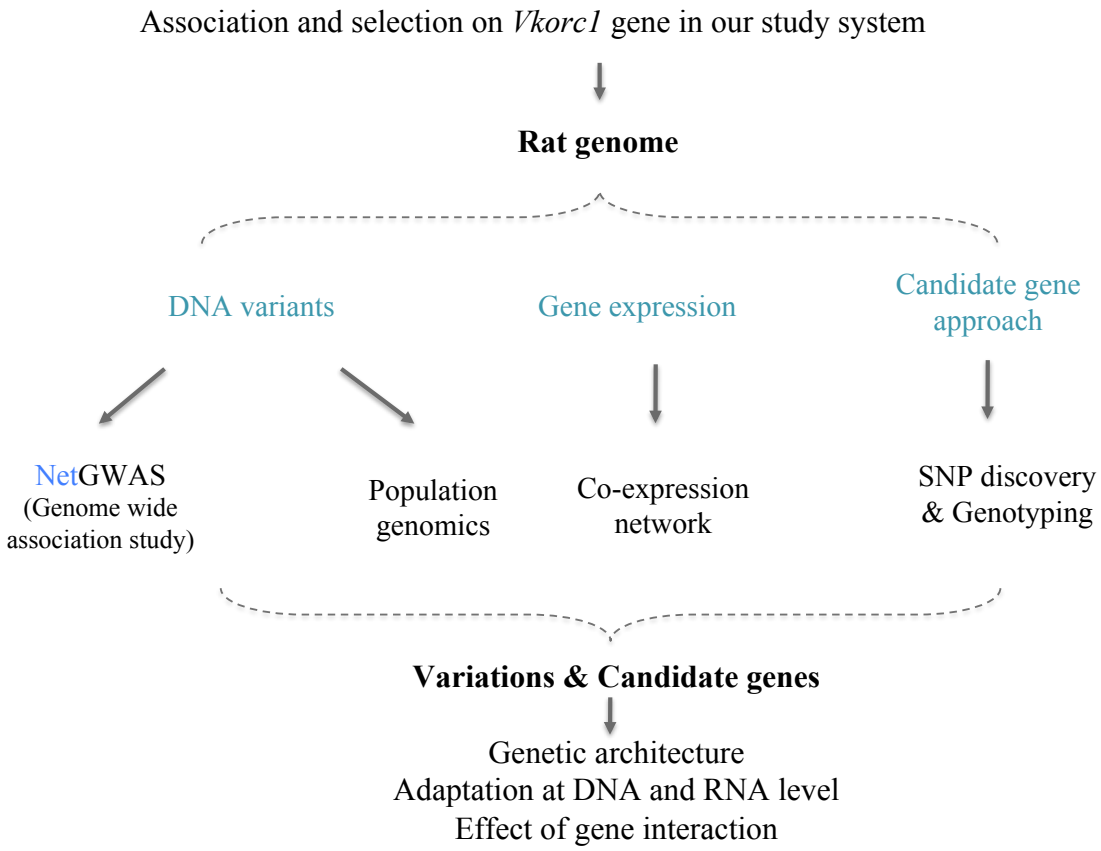


Figure 1.4 –Overview of workflow and major methological steps and innovations.

In addition to *Vkorc1*, I identified other candidate genes related to warfarin using network-guided genomic analyses based on variation data (SNP array) (Chapter 3) and microarray expression data (Chapter 4). Beyond the conventional GWAS and expression analysis of identifying candidate genes along the genome, we incorporated the information of gene interactions, which facilitated the candidate gene identification, and revealed relationships among multiple genes. The candidate gene list would be of interest to the study of warfarin therapy and genetic forecasting of dose in personalized medication. In this dissertation, not only we learned the genetic architecture of warfarin resistance, I also constructed a gene interaction network map for multiple candidate genes to understand their functional relevance, which can hardly be learned in traditional GWAS and expression analysis.

It connects evolutionary biology and medical genetic research, and has important implication in public health and agriculture. The developed algorithms and tools would be generally applied to other systems in cancer or disease genetic research. A key prediction of the thesis is that knowledge, or inferences, of gene interactions enable a systematic dissection of the genetics of adaptive traits, as I predict that a significant fraction of additional genes involved in warfarin resistance can be connected to the main resistance gene in terms of function.

Chapter 2

Balancing selection on an overdominant *de novo Vkorc1*-Y139C mutation in warfarin resistant Norway rat populations

Abstract

Field studies conducted in the 1977 have revealed that warfarin rodenticide resistance in Norway rats (*Rattus novergicus*) is an example of an overdominant mutation under balancing selection. However, the selection mode at the locus has not been studied with molecular population genetics; except that in the early 2000s analyses of microsatellite marker polymorphisms in free-living rat populations revealed the position of the resistance gene in form of localized and elevated levels of linkage disequilibrium, reduced polymorphisms, and genetic differentiation between populations of different resistance levels. These studies all pre-dated the time of the molecular identification of the warfarin resistance gene; now known to be encoded by a vitamin K epoxide reducatse subcomponent 1 (*Vkorc1*). Moreover, specific mutations in *Vkorc1* have now been shown to confer resistance during in vivo and in vitro studies. Foremost a mutation altering a Y at position 139 of the amino-acid sequence to a C is known to encode resistance to warfarin and to some of its derivatives, including bromadiolone and difeancoum. Here we conduct an analysis of single-nucleotide polymorphism (SNP) variation in Norway rats from Germany where the Y139C mutation is highly prevalent (>70%). From SNP typing on time-series samples of rats we estimate the selection coefficient at *Vkorc1* as $s = 0.3$. In conjunction with forward time simulations we show that observed patterns of polymorphism are consistent with the over-dominance model. We dated the mutation 'age' by comparing models of selective sweeps associated with *de novo* mutations with a starting frequency of $1/N_e$ with models of sweeps associated standing variants

of varying starting frequencies. We found evidence for a de novo origin of Y139C in *Vkorc1* in our study area, which could be interpreted as a *de novo* mutation that occurred or the arrival of an Y139C migrant rats in the area. During simulations we identified the setting of initial allele frequencies of neutral sites affect expected genetic variation patterns in a crucial fashion. Warfarin resistance in the Norway rat, as mediated by an Y139C mutation in *Vkorc1*, thus is confirmed as an example of an overdominant mutation under balancing selection.

2.1. Introduction

A fundamental challenge in evolutionary biology is to understand the genetic basis of adaptation. Amongst the many crucial steps involved in the dissection of adaptive traits is the identification of the genes and specific variants mapping within or around them that contribute to the trait (Barrett and Hoekstra 2011). Moroever, documentation of the relative adaptive contribution of the variants is as essential as documenting the adaptive value of the trait (Barrett and Hoekstra 2011).

Theoretical and molecular population genetic approaches thus have been developed that aid the detection and characterization of genes and specific variants in terms of their contributions to adaptation. Specifically, positive selection usually drives beneficial alleles from low to high frequency, or as it is generally modeled, to fixation (Pritchard and Di Rienzo 2010). The process affects physically linked neutral sites in that these rise in frequency alongside the selected site, and the region around a selected site will be ‘swept’ clean of genetic polymorphism. Thus, selected regions

can be detected in form of reduced genetic variation and extended linkage disequilibrium structure (Kim and Stephan 2003; Kim and Nielsen 2004; Stephan, Song, and Langley 2006). This effect, termed 'selective sweep' by 'genetic hitchhiking' (Smith and Haigh 1974) needs to be tested rigorous ly against neutral models that consider genetic drift, recombination and population structure.

However, while the approach has been applied to map the warfarin resistance gene by using patterns of polymorphism at neutral microsatellite loci (Kohn, Pelz, and Wayne 2000), the results were ambiguous in terms of the details regarding the selective regime that caused the observed patterns. In addition, at the time the resistance gene was not known and genome sequences and genomic tools to assay variation were unavailable, and thus, the study, results and analyses conducted at that time were necessarily limited when compared to the current state-of-the–art such analyses. In fact, previous analysis of microsatellite data appeared to contradict the expectations in that directional selection, rather than balancing selection on an overdominant warfarin resistance variant, could explain the observed patterns of variation (Kohn, Pelz, and Wayne 2000). In addition and crucially, unlike in 2000 the warfarin resistance gene is now known ((Li et al. 2004; Rost et al. 2004; Pelz et al. 2005); see below) such that the analyses can be targeted at the known variants in the gene and the directly adjacent neutral sites.

Warfarin works by inhibiting the activity of the vitamin K 2,3-epoxide reductase complex (VKOR) encoded *Vkorc1*, and subsequently impairing the vitamin K-dependent g-carboxylation system, which is essential for the activation of blood

coagulation factors and other vitamin K-dependent proteins (c.f. Appendix 8, Figure 5.2) (Presnell and Stafford 2002; Stafford 2005). Thus susceptible rats ingesting the poison succumb due to internal hemorrhaging while both homozygous resistant and heterozygous rats are resistant to the rodenticide.

As the first-generation rodent poison used since 1950s, warfarin (a derivative of 4-hydroxycoumarin) has imposed intense selection pressure on rat (*Rattus norvegicus*) populations (Hans-Joachim, Detlef, and Gerhard 1995; Kohn, Pelz, and Wayne 2000). However, a mere 8-10 years later, resistance towards warfarin has become prevalent in Germany (1967) and other European coutries (Boyle 1960; Lund 1964; population and agriculture 1986). The molecular basis for this resistance has been identified as the Y139C mutation in *Vkorc1* (vitamin K epoxide reductase complex, subunit 1) (Rost et al. 2004; Pelz et al. 2005). The Y->C amino acid change at position 139 (A to G mutation in *Vkorc1* exon 3) leads to a ~ 42% reduction of the basal *in vitro* VKOR activity but protects VKOR activity in the presence of warfarin at ~ 20%, which appears to be sufficient to maintain blood coagulation at levels that do not impair fitness (Pelz et al. 2005; Rost et al. 2009). The Y139C mutation has spread through a vast geographic area in northwestern Germany in response to warfarin and other anticoagulants (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003)). In this resistance area the resistance phenotype was determined for hundreds of rats by using a blood clotting response (BCR) test (Kohn, Pelz, and Wayne 2003).

However, according to the classical literature (Partridge 1979; Partridge 1980) a pleiotropic fitness cost in form of impaired blood coagulation is imposed on the homozygous resistant genotype. Recently, additional evidence has been collected showing that homozygous resistant rats display a more complex set of traits that likely are detrimental under intense competition in the wild. First, it has been recognized that resistant rats have reduced VKOR enzyme activity (Pelz et al. 2005; Rost et al. 2009) and higher requirement for vitamin K to maintain normal blood-clotting function in the absence of warfarin (Partridge 1979; Markussen et al. 2003). This results in the homozygous resistant rats having longer blood clotting times (Jacob et al. 2012). Second, in a *Vkorc1* mutant laboratory strain where Y139C has been introgressed experimentally a reduction in offspring numbers was reported (Jacob et al. 2012). Third, reduced growth rate have been observed in resistant rats from Welsh (Smith, Townsend, and Smith 1991) but the opposite was reported also (Smith et al. 1993). Vitamin K deficiency is expected to result in lower activities of vitamin K dependent proteins, notable here Matrix Gla protein (MGP), which is a vascular calcification inhibitor (Danziger 2008; Kohn, Price, and Pelz 2008).

Warfarin resistance has been a classical example of an adaptive overdominant locus under balancing selection since its initial discoveries (Greaves et al. 1977). For the warfarin resistance locus in wild rats early estimations of the relative fitness ratios based on the sampling of rats during fieldwork provided a fitness model for the three phenotypic classes and genotypes (Table 2.1). However, this classical textbook example for an overdominant locus has not been revisited with state-of-the art molecular population genetics and rigorous computer simulations.

Table 2.1 – Fitness models at the warfarin resistance locus

		Genotype		
		S/S	S/R	R/R
Fitness model		1-*s*	1-ds	1-*t*
No selection		1	1	1
Directional selection	Recessive	1-*s*	1-*s*	1
	Codominance	1-*s*	1-1/2*s*	1
	Dominance	1-*s* (0.70)	1 (1.00)	1 (1.00)
Balancing selection	Over-dominance	1-*s* (0.70)	1 (1.00)	1-*t* (0.90)
	(Greaves et al. 1977)	0.68	1.00	0.37
Penetrance model		0.07	0.95	0.98

S - wild type susceptible allele, R - resistance allele (Y139C).
s - selection coefficient for S/S genotype.
t - fitness cost of R/R.
d - dominance parameter.
Values in the parentheses denote the fitness ratio estimated by our study; the estimation from Greaves et al. is shown separately (Greaves et al. 1977).
The penetrance model for the resistance mutation in *Vkorc1* gene is calculated and averaged across populations (see methods).

Warfarin resistance appeared rapidly after the introduction of warfarin as a rodenticide in the 1950s. This raises the question of broad evolutionary biological relevance whether adaptive variants often are selected from pre-existing standing genetic variants rather than by selection on *de novo* mutations. Theory predicts that selection on the former class of variants results in narrower valleys of reduced polymorphism when compared to such valleys associated with new beneficial mutations (Figure 3 in Barrett et al.) (Barrett and Schluter 2008). This is explained by the expectation that standing variations have longer histories preceding selection during which recombination and recurrent mutation add polymorphism at linked

neutral sites. In other words, adaptive standing variants are associated with different haplotypes as opposed to one haplotype (Barrett and Schluter 2008). Others have speculated that resistance has evolved from de novo mutation (Pelz et al. 2005) or from standing variation (Barrett and Schluter 2008).

Here we use molecular population genetic (single nucleotide polymorphisms, SNPs) data collected in a large sample of wild-caught rats for which the resistance phenotype and the *Vkorc1* genotypes are known (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). We formally analyze the mode, timing and strength of selection on warfarin resistance as mediated by Y139C in *Vkorc1*.

2.2. Methods and Materials

2.2.1. Sampling

We chose 29 rats of a larger population of wild rats where warfarin and bromadiolone resistance segregates (designated as NW sample) that were maintained in the laboratory by Hans-Joachim Pelz at the Federal Research Institute for Cultivated Plants, Julius Kuehn Institute (JKI), in Muenster, Germany. These rats were caught in the wild from a highly resistant population in northwestern Germany near Muenster, Germany, and located within a vast geographic region in northwestern Germany where resistance is highly prevalent. From this resistance area we previously obtained samples of wild rats from 35 farms during a three years' project investigating the resistance to rodenticides including warfarin (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). Here, based on DNA quality of these we

included 728 rats from this previous study for further analyses. Furthermore, we chose 12 non-resistant rats from a fully susceptible population (designated as sample LH) sampled previously about 300 km away from the resistant area (Kohn, Pelz, and Wayne 2000).

In the resistant area, populations with sample sizes > 10 were kept for further analysis; resulting in a sample of 668 rats collected on 17 farms (Appendix 1). Of these 14 populations were exposed to warfarin and later treated with other anticoagulants rodenticides. We used the abbreviation names for 5 populations (WU-pop24, KB-pop11, TH-pop23&16, SP-pop20, LH-pop4f) as mentioned in a previous study (Kohn, Pelz, and Wayne 2000). We labeled other populations by their farm numbers adopted from a previous study that utilized these samples (Kohn, Pelz, and Wayne 2003).

Resistance to rodenticides was physiologically determined with a blood clotting response (BCR) test (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003) and were available for 599 of the 668 rats. The resistance level, R%, for each population is calculated using the number of resistant individuals divided by the number of samples with resistance phenotype assigned. Note that heterozygous rats (R/S) and homozygous rats (R/R) are resistant.

Our data have a temporal dimension in that careful monitoring in the field documented the incidence of resistance frequencies during controlled field experimentations where populations were sampled before and after treatments with the anticoagulants in the field from 1996 to 1998. Thus, we were able to take

advantage of time-series samples to estimate the selection coefficient on *Vkorc1* Y139C, as we have collected the genotypes at *Vkorc1* for rats in this study (see below).

2.2.2. Phylogeny of the resistance gene *Vkorc1*

Pseudogenes of *Vkorc1* exist in mammalian genomes, and thus, we examined the currently annotated *Vkorc1* of the rat in a phylogenetic framework to establish orthology to the human gene and to characterize briefly the pseudogenes that occur in the rat genome. We constructed the phylogenetic tree of the known resistance gene *Vkorc1* (vitamin K epoxide reductase complex, subunit 1) and its paralog *Vkorc1l1* (*Vkorc1-like protein 1*) among 6 species: human (*Homo sapiens*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), dog (*Canis lupus familiaris*), Cattle (*Bos taurus*), Fugu (*Takifugu rubripes*) (Figure 2.1). Multiple sequence alignments of genomic sequences were conducted in ClustalW 2.0.10 at default settings (Larkin et al. 2007). Phylogenetic trees of nucleotide and protein sequences were estimated using Bayesian inference (BI) with fugu as outgroup. The Bayesian phylogenetic trees were estimated using the MrBayes 3.1.2 program (Ronquist and Huelsenbeck 2003) with two simultaneous Markov Chain Monte Carlo (MCMC) chains run for 1,000,000 generations and sample frequency of every 100 generations with burn-in = 2,500. The Generalized Time Reversible evolutionary model (GTR; Nst=6) was applied.

2.2.3. Genotype data of *Vkorc1* and the surrounding region

The Y139C mutation on the *Vkorc1* gene has been identified as the main mutation that confers warfarin resistance in Norway rats from the study area (Pelz et al. 2005). We genotyped all samples from the NW population and 691 samples from 19 other populations by using High Resolution Melting (HRM) SNP analysis (Wittwer et al. 2003). The Type-it HRM PCR Kit was purchased from QIAGEN (http://www.qiagen.com, a provider of sample and assay technologies). The primers covering the identified SNP underlying Y139C targeted with HRM genotyping (Appendix 2) were designed in Primer3 (http://frodo.wi.mit.edu/primer3, access May 2011). For each sample, 0.65 ul genomic DNA was added to reaction volume of 10 ul, which includes 5 ul HRM PCR Master Mix, 2.95 ul RNase-free water and 1.4 ul primer Mix.

HRM analyses were conducted on the Rotor-Gene Q (QIAGEN's real-time PCR cycler), with an initial PCR activation of 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 10 seconds, annealing and extension at 55°C for 30 seconds; the following HRM were performed in a temperature range from 75°C-90°C with 0.1°C increments at each step. The genotypes for each rat sample were identified with the threshold set at a 'confidence percentage' > 85% (White, Hall, and Cross 2007) in the HRM analysis that compared with 9 reference samples with the *Vkorc1* genotypes (GG, GA, AA) as determined by DNA Sanger sequencing.

For the samples of LH (non-resistant population) and NW (wild-derived laboratory resistant rats), the genotype data of 100 SNPs flanking *Vkorc1* (from 172

Mb to 202 Mb on chromosome 1) were collected. Eighty of these SNPs were assayed on the Affymetrix Rat 10k SNP array platform. We refer to these SNPs by their assay ID provided along with the Affymetrix array and preceded by a 'S' (Appendix 3). Data on 19 SNPs that map within 2 Mb of *Vkorc1* for LH and NW rats were obtained by PCR amplification and Sanger sequencing in our laboratory (unpublished). These SNPs were named by their gene names and their physical position on Chromosome 1.

2.2.4. Genotype-phenotype association test

Although there is a broad consensus that *Vkorc1* is a main resistance factor, surprisingly, this has not been tested formally yet at the level of populations using association studies accounting for various factors confounding association analyses. Amongst these are population structure and sample stratification, and the effect of genetic hitching on false positive association of linked neutral sites. We computed genetic association of the Y139C mutation in *Vkorc1* with the warfarin resistance phenotype as determined by BCR test for rats from 14 natural populations (Cochran-Mantel-Haenszel tests controlling for sex; Table 2.3). Tests were done as implemented in PLINK v1.07 (Purcell et al. 2007).

For the NW population association tests using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) were run for all 100 SNPs, including those SNPs that map within the ~30 Mb 5' and 3' flanking regions of *Vkorc1* (Purcell et al. 2007). Association tests were run assuming a simple allelic model with 1 degree-of-freedom. To control for the effect of sex the CMH (Cochran–Mantel–

Haenszel) test was applied. Raw P-values and P-values adjusted using Holm's (1979) step-down multiple correction method were calculated (Appendix 3). The Bonferroni correction method were similar to those obtained by using Holm's correction (not shown). We plotted the $-\log_{10}$ of the P-values along the 30 Mb region on chromosome 1 that carries the Y139C mutation in *Vkorc1* (Figure 2.2A).

2.2.5. Imputation-based association test between genotype and phenotype

In addition to traditional association tests we implemented an imputation-based Bayesian regression association method implemented in the software package BimBam (Scheet and Stephens 2006; Servin and Stephens 2007; Guan and Stephens 2008). Using the prior D2 from (Servin and Stephens 2007) we computed Bayes Factors (BF), which measure the strength of association, and at the same time obtained P-values by 1,000 random permutations. By definition, the BF for the null hypothesis of no association between SNP and trait value, H_0, is set as 1. Within this framework a $\log_{10}$(BF) of 2 indicates that the marker-trait association is 100 times more likely under the prior model than under H_0. For each SNP we computed three BFs each assuming an additive model (BF1), a dominance model (BF2), and an over-dominance model (BF3). We conducted these calculations by adjusting the values of the parameters σ_a (additive effect) and σ_d (dominance effect). BF1 was calculated under additive model by averaging over $\sigma_a = 0.1, 0.2, 0.4$ and $\sigma_d = \sigma_a /4$ as previously was suggested (Servin and Stephens 2007; Stephens and Balding 2009). BF2 was calculated by averaging over $\sigma_a = 0.1, 0.2, 0.4$ and $\sigma_d = \sigma_a$ to increase the weight on the dominance model. In addition, considering the over-dominance model, we

computed BF3 using $\sigma_a = 0.45$, $\sigma_d = 0.55$. The $\log_{10}$ values obtained for the BFs and the corresponding P-values are shown in Figure 2.2A (c.f. Appendix 3).

Furthermore, we calculated the posterior probabilities of association (PPA) to evaluate how many SNPs are truly associated with the phenotype (Stephens and Balding 2009). We used a prior distribution of up to 2 causal SNPs and gave weights to the ratios 2/1/0 to 1-SNP, 2-SNP and >2-SNP models to calculate the PPA by averaging over the 30 Mb region, but also BY ANALYZING each SNP SEPARATELY or BY AVERAGING OVER each 2-SNP pair. Models containing the same number of causal SNPs were assumed as equal priors that were weighted by $\binom{n}{l}$; where n is the number of SNPs mapping and assayed in the region (here 100) and *l* denotes the number of causal SNPs. We plotted the marginal PPA for each SNP in the region using the statistical software package R to assess whether there is any significance evidence (PPA > 0.5) for causal SNPs other than Y139C *Vkorc1* (Figure 2.3).

2.2.6. Linkage disequilibrium (LD) around *Vkorc1*

For the non-resistant population LH and resistant NW rats, linkage disequilibrium (LD) in the 30 Mb region was calculated in PLINK (Purcell et al. 2007) using the measure of r-square for phased data, which represents the statistical correlation of allele frequencies at any two loci. A heat map of LD (r-square) matrix for the NW population (Figure 2.2A) and the LH population (Figure 2.2B) were plotted across the 30 Mb region using R.

2.2.7. Estimation of the selection coefficient at *Vkorc1*

To quantify the selection coefficient on the Y139C mutation in *Vkorc1* we jointly estimated *s* and the effective population size N_e for *Vkorc1* based on time-series sampling data of allele frequencies using an expanded hidden Markov model (HMM) (Bollback, York, and Nielsen 2008). This method assumed that the change of allele frequency was primarily driven by selection on a co-dominant locus; i.e. assuming *s* heterozygotes and *2s* for homozygotes.

For each population, we binned the samples into several time-points based on the sampling time and the rodenticides treatment. Data points with only one rat sample per time point were removed from the analysis. Populations were exposed to warfarin first and subsequently were exposed to a second-generation anticoagulant bromadiolone (and select populations were exposed to difenacoum and/or brodifacoum). However, here we only considered the first set of data points where populations were sampled prior to warfarin use for rodent control and then after the exposure to warfarin. The Y139C allele frequencies prior and after warfarin exposure are provided in Table 2.3.

With the assumption of 3 months a generation (Adams and Boice 1983), the numerical HMM integration was performed for *s* ranging from -0.5 to 5, and N_e ranging from 100 to 100,000, with a grid size >= 500 as recommended. We used a grid size of 1,000. The maximum likelihood estimates converged towards $N_e = 10^3$. Thus, we performed the profile likelihood analysis for the parameter settings $s = -1$ to 1 and N_e=1,000, to *s* for each population separately. Corresponding 95% confidence

intervals for *s* were estimated based on the chi-square distribution. Assuming N_e=10,000, we obtained similar estimates of *s* (not shown). For some populations we were unable to estimate *s* because no samples with the *Vkorc1* mutation were available at the end of sampling period (Table 2.3). It is known that both the Y139C heterozygote and homozygote genotypes are warfarin resistant, i.e. the mutation is penetrance (Table 2.1) (Pelz et al. 2005; Kohn, Price, and Pelz 2008). Hence, the fitness values for the susceptible homozygote, the Y139Y/C heterozygote and the Y139C homozygote are set as 1, 1 + *s* and 1 + 2*s*, respectively, to reflect co-dominance. Consequently, *s* of homozygous susceptible rats (S/S) was corrected by 1 - 1 / (1 + 2*s*). The average *s* calculated across populations was then calculated (Table 2.3) and used in our forward population genetic simulations (see below).

2.2.8. Population genetic analysis

2.2.8.1. Recombination rate

We estimated the recombination rate using PHASE 2.1.1 (Stephens and Li 2003; Crawford et al. 2004) for the *Vkorc1* gene flanking region, i.e. based on polymorphism data collected for 100 SNPs mapping between chromosome locations 172 Mb and 202 Mb on rat chromosome 1. The analysis was done for the sample taken from population LH (non-resistant) a sample taken from population NW. The parameter settings underlying this analysis of recombination were as follows: Number of iterations=100; thinning interval=1; burn-in=100. The default recombination model was adopted and analyzed assuming a prior of the population recombination rate rho = $4N_e\mathrm{c} = 2.4\times10^{-5}$ (Jensen-Seaman et al. 2004)). Using the default prior of

recombination rate as rho = 4×10^{-4} yielded similar results (not shown). As recommended, for each input file, we ran the algorithm implemented in the PHASE software 5 times to select results with the highest value for the goodness of fit estimation. Each final run was set to be 10 times longer than other iterations to obtain a better estimation of the recombination rate. We calculated the median of rho and the median of varying factor lamda (the rate between locus *i*-1 and locus *i* exceeds the background rate given in the first column of the recombination estimation file) for each SNP interval (distance between two SNPs). The recombination rate associated with each SNP interval was calculated as rho x lamda, and the average recombination rate was obtained for N_e = 1,000. We estimated r = 0.002 per megabase for the sample taken from the susceptible population LH, and in our simulations and analyses was taken as the background recombination rate and not confounded by the effect of selection on polymorphism (e.g. by reducing N_e).

As described in detail below, we conducted forward time population genetic simulations for ideal populations and assuming sets of population variation parameters, such as allele frequencies prior to selection, assuming a constant recombination rate r = 0.006 per megabase per generation as estimated for Norway rats laboratory strain to show a general trend (Jensen-Seaman et al. 2004). For the forward simulations that were run using parameter setting, again, e.g. allele frequencies prior to warfarin selection, taken from the empirical data as taken from the susceptible population LH and the recombination rate r = 0.002 per megabase as estimated from these empirical allele frequency data to compare with the observation data.

2.2.8.2. *Vkorc1* haplotype structure

To infer which allele of each *Vkorc1*-linked SNP initially was associated with Y139C when the mutation was selected first we used the haplotype reconstruction software PHASE 2.1.1, which implements a Bayesian statistical approach and EM (Expectation Maximization) algorithm (Stephens, Smith, and Donnelly 2001; Stephens and Scheet 2005). We inferred the haplotypes for pairs of SNPs; one linked SNP and the SNP underlying the Y139C mutation in *Vkorc1*. We allowed a recombination model with a suggested recombination rate 0.00001 as prior, and performed the permutation of case-control (resistant and non-resistant) test to increase the inferring power. As above, we ran PHASE 5 times with each final run 10 times longer to estimate the likely haplotypes. The probability of any linked SNP allele to be on the same haplotype as the SNP underlying Y139C was deduced from the inferred haplotype frequencies; i.e. the most commonly inferred haplotypes were taken as those representing the original haplotype the introduced Y139C mediated resistance in the population.

2.2.8.3. Allele frequencies of *Vkorc1* and linked sites

For each rat the time point of sampling was recorded such that for our sample it is known whether any rat was sampled before warfarin rodenticide application to the population or thereafter. Thus, the allele frequency of the resistance mutation Y139C in *Vkorc1* gene could be computed for these time points in each of the 19 natural populations of rats that underwent experimental rodent control with warfarin. The allele frequencies prior and after warfarin treatments are shown in Table 2.3. The

allele frequency change over time is shown for 7 populations where high resistance levels (R% > 80%) were encountered (Figure 2.4); populations with low R% were not shown since they may not be in the comparable stage or level of evolutionary history in terms of warfain selection. The sample taken from the NW population consisted out of 20 resistant and 9 non-resistant rats (Appendix 1), which were derived from wild rats sampled from a highly resistant population with estimated resistance frequency of ~90% as estimated in the field. To recover the population statistics of the original (unsampled) population, we performed the permutation of NW samples according to three resistant levels: 85%, 90% and 95% respectively and calculated the average allele frequency and other population parameters for the NW population. For further analysis, we used the population statistics obtained through the permutations obtained from re-sampling form a population with a resistance frequency of 90%.

Inferred haplotype frequencies were used to calculate the allele frequency of the SNP alleles that prior to warfarin selection were on the same haplotype as the resistance mutation. These inferences were done based on the data collected for the non-resistant population LH and the resistant population NW (recorded in Appendix 3).

2.2.8.4. Polymorphism measures

We calculated two polymorphism measures at each SNP in each population for both the empirical and simulated data (see below):

Nucleotide diversity θ_π was estimated using: $$\theta_\pi = 1 - \frac{\sum_i \binom{n_i}{2}}{\binom{n}{2}}$$

Equation 2.1 – Nucleotide diversity.

where n_i is the count of allele i at each SNP in the population, and $n = \sum n_i$ (Hohenlohe et al. 2010).

Observed heterozygosity H_{obs} was obtained by calculating the proportion of individuals with heterozygote genotypes at each SNP. Averages for both statistics were calculated across SNPs.

2.2.9. Fitness model definitions at *Vkorc1* and forward-time simulations

Forward time simulations were done using Python scripts (provided upon request) designed to implement algorithms provided by the environment of SimuPOP (Peng and Kimmel 2005; Peng, Amos, and Kimmel 2007). Two types of simulations are distinguished throughout. First, simulations starting from simplified population samples where, e.g. we set starting allele frequencies. We refer to the as simulations based on assumed data. Second, we ran simulations starting from allele frequencies taken from the observed data prior to selection, which we chose from the susceptible population LH. We refer to these as simulations based on empirical data. We assumed 1,000 diploid individuals in each simulated population but similar results were obtained with a population size of 10,000 (data not shown). For each forward-time simulation we ran 1,000 replicates, and the distributions of results were used to generate later statistical summaries including averages and confidence levels.

Furthermore, we assumed random mating and an estimated recombination rate: $r = 6 \times 10^{-3}$ per megabase per generation for simulations that starting with assumed allele frequencies to show the general trend. We assumed $r = 2 \times 10^{-3}$ per megabase per generation for the simulations that built on the empirical data for comparing with observed genetic variation in empirical data (Jensen-Seaman et al. 2004).

We simulated the beginning of selection with warfarin in 1950s (generation 0). We could consider 45 years from 1953 (the first time warfarin was used in Europe) to 1998 (the end of our sampling date), which corresponds to ~31 years with the knowledge that warfarin resistant rats have been noticed first in Germany in 1967 (population and agriculture 1986). These time periods ALL corresponded to at most 200 generations when assuming 4 generations a year, and to at least 93 generations when assuming 3 generations a year (Adams and Boice 1983). In practice, 200 generation was used for our simulations of hypothetical populations to show the general trend until 200 generations; 150 generations were used for simulation done based on empirical data as the middle point between 100 and 200 generations (Table 2.2).

Table 2.2 – Simulation settings.

Simulation Type	N	Time frame	SNPs	Simulation Models	
				Mutation age	Fitness model
					Dominance & New Dominance & Standing	Overdominance & New Overdominance & Standing
Assumed data	10^3	200 gens	301 sites across 30 Mb region (0.1 Mb density) & equal initial allelefreq			

						Neutral	
Empirical data	10^3	150 gens	100 SNPs across 30 Mb around *Vkorc1* with initial allelefreq from LH (nonresistant)	Mutation age	Fitness model		
					Dominance & New	Overdominance & New	
					Dominance & Standing	Overdominance & Standing	

N: population size. Fitness model (c.f. Table 2.1).

2.2.9.1. Simulation based on assumed data

We structured the data such that 301 neutral SNP sites were evenly distributed along a 30 Mb region on chromosome 1 (density = 0.1 Mb), with the 151th SNP localized I the center and under selection. We assigned the same initial allele frequency to the alleles at each SNPs to range from 0.1 to 0.9.

2.2.9.2. Simulations based on empirical data

Following the structure of our observed data we consider 100 SNPs located in a 30 Mb region flanking *Vkorc1* on both sides. Their initial allele frequencies were set according to the values observed in the fully susceptible population LH population.

2.2.9.3. Standing variation or new mutation

We considered two scenarios that differ in terms of the origin of the Y139C mutation. First, we assume that the allele is a *de novo* mutation starting with frequency $1/2N_e$ in generation one.

Second we consider a scenario where the Y139C mutation pre-dated the application of warfarin, and thus was present in the population at some frequency

higher than $1/2N_e$ at generation one. We attempted to simulate this scenario using a few additional pieces of information available for Norway rats and their colonization history of Europe. It is believed that *R. norvegicus* colonized Europe ~1716-1796 (Pelz et al. 2005). We use the simplified time window spanning the years from 1750s-1950s, which correspond to 600-800 generations in rats assuming 3-4 generations per year. We then simulated the situation that the Y139C mutation occurred about 250 years ago *de novo*, i.e. we assume that the mutation was not present in the founding populations of rats that came from Asia, but instead, that the mutation is comparatively recent but pre-dates the introduction of warfarin. We assumed this scenario such that a reasonably wide range of standing allele frequencies can be generated. Evolution then was simulated to occur from generations -700 to generation zero under the neutral model and then under selection since the use of warfarin in 1950s (generation zero). Selection was model considering the various models and parameter settings for directional selection and balancing selection.

Specifically, the fitness model under selection is described as the first row in Table 2.1. There, S represents the wild type allele of the SNP under selection, and R represents the mutation. The selection coefficient s is set to act to select against non-resistant rats, and t is the fitness cost that the Y139C mutation incurs on the homozygous mutant genotype. Here we only considered the dominance and over-dominance models since rats with both genotypes S/R and R/R are resistant to warfarin (Penetrance model in Table 2.1) (Pelz et al. 2005; Kohn, Price, and Pelz 2008). Finally, drift only was simulated and compared to the observed results in both the populations LH (all susceptible) and NW (~90% resistance frequency).

The selection coefficient $s = 0.3$ has been estimated as described above based on the time-series sampling of rats. In the directional selection model, both the S/R and R/R genotypes have a fitness value of 1 assigned to them. Under a balancing selection model we estimated the fitness cost t of the R/R genotype by solving the equation $P_R = s / (s + t)$ (c.f. page 217 in (Hartl and Clark 2007) for t; where P_R is the equilibrium allele frequency of the mutated allele. With a $P_R \sim 0.75$ estimated from the empirical data at *Vkorc1* accross our natural populations (Figure 2.4) and $s = 0.3$ (Table 2.3), we obtain $t = 0.1$ as the fitness cost for the mutant homozygotes (R/R).

Finally, the combination of the above two scenarios describing the evolution of resistance by a new mutation or from standing variation, and the 2 fitness models describing directional selection on a dominant mutation or balancing selection on a overdominant mutation, yielded a 2 x 2 table (Table 2.2).

2.2.9.4. Statistics from simulations

We obtained several population genetic statistics from the simulations and compared them with the empirical data:

For each generation and for each SNP we tabulated the allele frequencies of the allele (by inference or as per simulation settings, see above) that initially was linked to the mutation in *Vkorc1* on the same haplotype, and we tabulated R^2 and D' measuring LD between the selected SNP and linked neutral SNPs.

At the final generation we tabulated genotype frequencies, observed heterozygosities, and the frequencies of the haplotypes where the mutation in *Vkorc1* and the linked allele at SNPs remained on the same haplotype.

The genotype-phenotype association test was conducted after sampling 35 cases and 15 controls from a simulated 1,000 individuals. The penetrance model for the resistance mutation in *Vkorc1* was estimated by calculating the proportion of individuals carrying the genotype SS, SR and RR that are resistant to warfarin in the population. The averaged penetrance values across population NW and all other populations are given in Table 2.1.

2.2.9.5. Simulations considering a range of selection coefficients

We performed simulations where directional selection acts on a *de novo* mutation. The selection coefficients we examined are 0.01, 0.05, 0.1, 0.2, 0.3 and 0.5, and the populations evolved for 2000, 1000, 400, 300, 200 generations, respectively, when they reached fixation of the beneficial allele. After fixation of the beneficial alleles we tabulated and plotted the allele frequencies along the chromosome to evaluate the expected size of the region affected by the selective sweeps under various scenarios and settings.

2.3. Results

2.3.1. Description of resistant and non-resistant populations

This chapter focused on one non-resistant (LH) and one resistant (NW) population. The population NW assayed on the SNP array is derived from wild rats (*Rattus norvegicus*) were trapped from a highly resistant area in the northwestern Germany and then kept in the lab for about 12-17 months allowing breeding (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003; Kohn, Price, and Pelz 2008). From this population, termed NW population, 20 resistant and 9 non-resistant rats were assayed for SNP variation on the Affymetrix SNP arrays. To infer the original population features (i.e. assuming most of the population was not sampled) we permuted the NW population by shuffling samples according to a 90% resistance level (the estimated resistance frequency of the total population as per personal communication with Hans-Joachim Pelz, at the Federal Research Institute for Cultivated Plants, Julius Kuehn Institute (JKI), in Muenster, Germany). When we considered other resistance levels (85%-95%) results were similar (not shown). From population LH (located about 300km away from the resistant area) we assayed 12 non-resistant rats.

In addition, we also studied 668 rats from 17 natural populations from the resistant area (Appendix 1). 14 populations were exposed to warfarin as well as other anticoagulant rodenticides. Warfarin has been used as a rodenticide in the field since 1950s (Hans-Joachim, Detlef, and Gerhard 1995; Kohn, Pelz, and Wayne 2000). Samples of rats trapped in the field between 1996-1999 were sampled at different

time points, including time points that pre-dated the application of warfarin in these populations and at time points that post-dated the application of warfarin (Pelz, Hänisch2, and Lauenstein 1995; Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). Notable, this is not to imply that these populations were exposed to warfarin for the first time, but were exposed to warfarin by our field studies for the first time. The sample sizes obtained from these populations ranged from N=19 to 73.

2.3.2. The rat ortholog of human biomarker *Vkorc1* is associated with warfarin resistance

Vkorc1 gene causes warfarin resistance in rodents (Pelz et al. 2005; Rost et al. 2009), and in humans *VKORC1* is known to be the primary target of warfarin. Polymorphisms in *VKORC1* can be used to make drug dosage predictions (Rost et al. 2004; McDonald et al. 2009; Pautas et al. 2009; Takeuchi et al. 2009). The rat gene *Vkorc1l1* (*Vkorc1*-like protein 1) is paralogous to *Vkorc1* (Rost et al. 2004). To learn the evolutionary relationships of these genes and to ascertain that we study the only functional *Vkorc1* in the rat genome we constructed a phylogenetic tree based on nucleotide gene sequence and amino acid sequence in 6 species, namely human (*Homo sapiens*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), dog (*Canis lupus familiaris*), cattle (*Bos taurus*), and fugu (*Takifugu rubripes*) as outgroup (Figure 2.1).

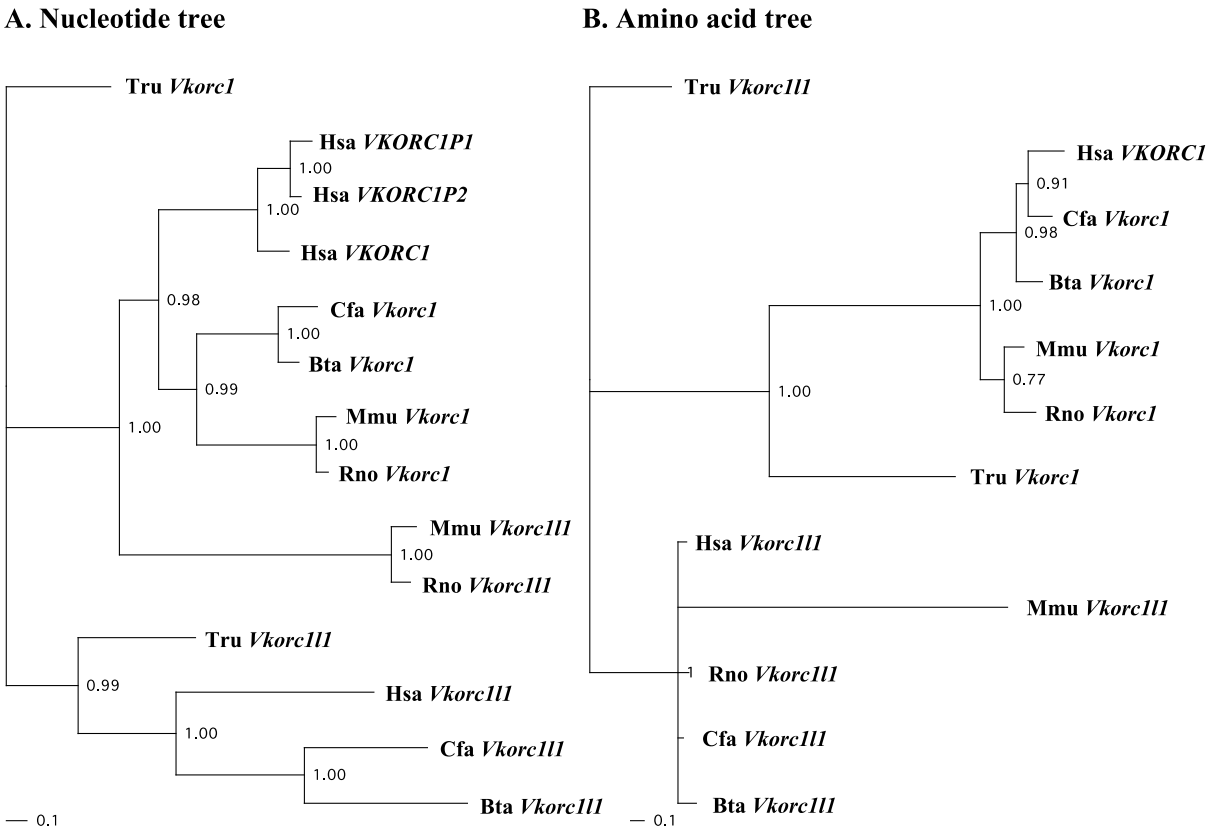


Figure 2.1 – Phylogenic relationships among *Vkorc1* and *Vkorc1l1*. A. DNA sequence phylogeny. B. Amino acid phylogeny. Bootstrap values are shown besides nodes. Scale bars indicate nucleotide and amino acid substitutions per site. Phylogenetic trees were estimated using Bayesian inference (BI) with two simultaneous Markov Chain Monte Carlo (MCMC) chains run for 1,000,000 generations and sampling of trees with a frequency of 1 every 100 generations with burn-in = 2,500. Hsa: human (*Homo sapiens*); Rno: rat (*Rattus norvegicus*); Mmu: mouse (*Mus musculus*), Cfa: dog (*Canis lupus familiaris*); Bta: cattle (*Bos taurus*); Tru: fugu (*Takifugu rubripes*).

The human *VKORC1* (4101 bp, 163 aa), grouped with two pseudogenes: *VKORC1P1* (chromosome X) and *VKORC1P2* (chromosome 1). It is the ortholog of

the rodent *Vkorc1* (2521 bp, 161 aa). It might be interesting to explore the potential function of pseudogenes in expression regulation (Long and Zhang 2012). Human *Vkorc1l1* is 81,543 bp, much longer than *Vkorc1* in terms of nucleotide sequence; but with 176 aa the pseudogene has an amino-acid sequence similar in length to *Vkorc1*. Rat *Vkorc1l1* also produce a 176 aa sequence but its nucleotide sequence length spans 46,299 bp. The complete history of gene duplication events remains ambiguous as the phylogeny based on the nucleotide sequences and based on amino-acid sequences differ (Figure 2.1). Nevertheless, the main intention here was to validate orthology between the well-studied human gene and the rat gene.

Numerous lines of investigations have identified *Vkorc1* as a major genetic factor underlying warfarin resistance in rodents as well as in humans. However, curiously, none of the studies have rigoriosly tested for association in wild rodent populations while controlling for demographics and genetic hitchhiking. In particular, previous studies tested for association in highly resistant rat populations where the extended haplotype due to hitchhiking leads to association between markers and resistance over vast genomic distances (Kohn, Pelz, and Wayne 2000).

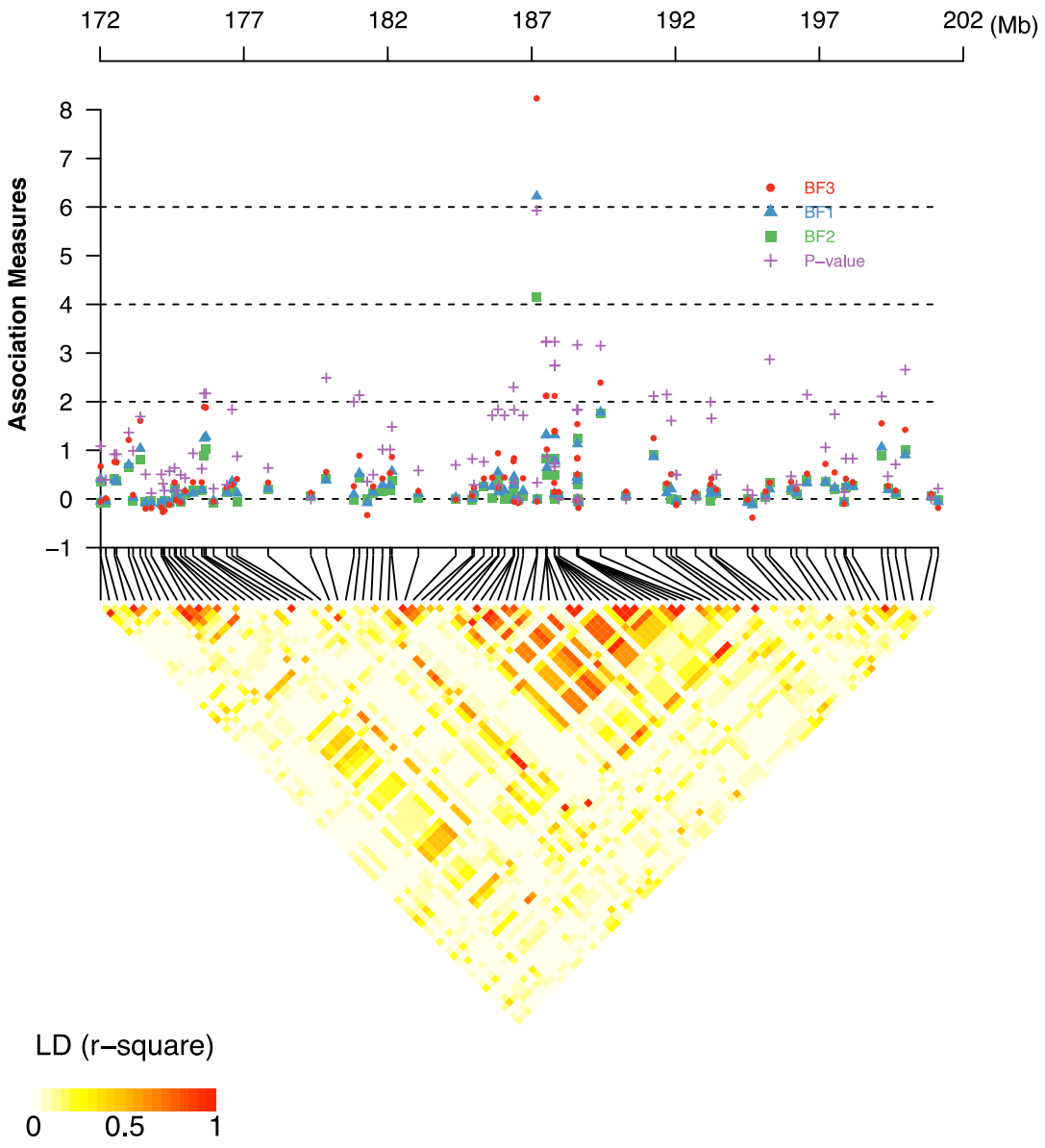
Here we tested for association between the resistance phenotype as determined by the BCR tests and the Y139C mutation, as well as 100 SNPs that map onto rat chromosome 1. For the population NW we find that *Vkorc1* is strongly associated with warfarin resistance (P-value = 1×10^{-6}) (Table 2.3). We also collected the genotype data of the Y139C mutation in *Vkorc1* gene and tested their genotype-phenotype associations in other 14 wild populations that range in resistance

frequencies from high to low. As shown in Table 2.3, *Vkorc1* is significantly (P-value < 0.05) associated with warfarin resistance in most (11 of 14) populations. Thus, here we reconfirmed that Y139C in *Vkorc1* gene is a major resistance factor underlying resistance.

2.3.3. Characterization of the selective sweep at *Vkorc1*

As suggested by a previous study (Kohn, Pelz, and Wayne 2000), we performed the association tests cross 30 Mb around *Vkorc1* (from 172 – 202 Mb on Chromosome 1). We observed association due to genetic hitchhiking of numerous linked SNPs with resistance and the Y139C mutation in population NW (Figure 2.2A). In fact, 40% SNPs within the region are significantly associated with the warfarin resistant/susceptible phenotypes. We computed three Bayes Factors (BF1-3) to measure the association strength of each SNP based on an imputation-based Bayesian approach (Servin and Stephens 2007) assuming either additive (BF1), dominance (BF2) and over-dominance (BF3). *Vkorc1* gene has the highest BF under the over-dominance model ($\log_{10}$BF3 = 8.235) (Figure 2.2A). Other than *Vkorc1*, no other SNP in the region has a $\log_{10}$BFs > 3, which suggests overdominance model has relatively strong power for detecting association of *Vkorc1* variant. 43% of the SNPs in the region have P-values < 0.05 for BF3, contrasting with 18% SNPs with P-values < 0.05 for BF3 on the whole Chromosome 1.

A.


172
177
182
187
192
197
202 (Mb)
Association Measures
8
7
6
5
4
3
2
1
0
−1
BF3
BF1
BF2
P−value
LD (r−square)
0
0.5
1

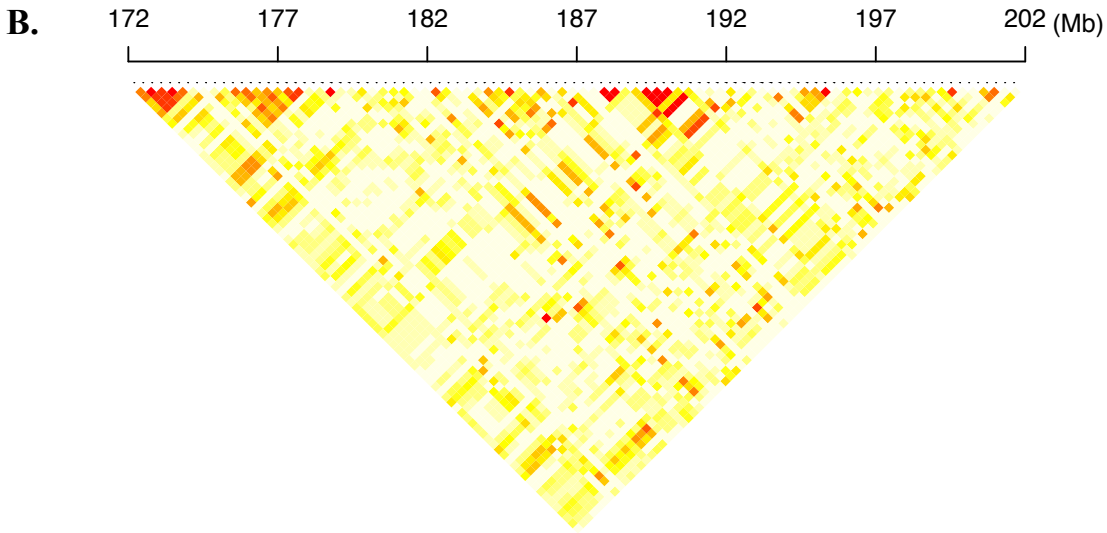


Figure 2.2 – Marker-trait association and linkage disequilibrium (LD) along rat chromosome 1. (A) Results are shown for the highly resistant (~90% warfarin resistance frequency) population NW. (B) The LD block of the same region for the non-resistant LH population.

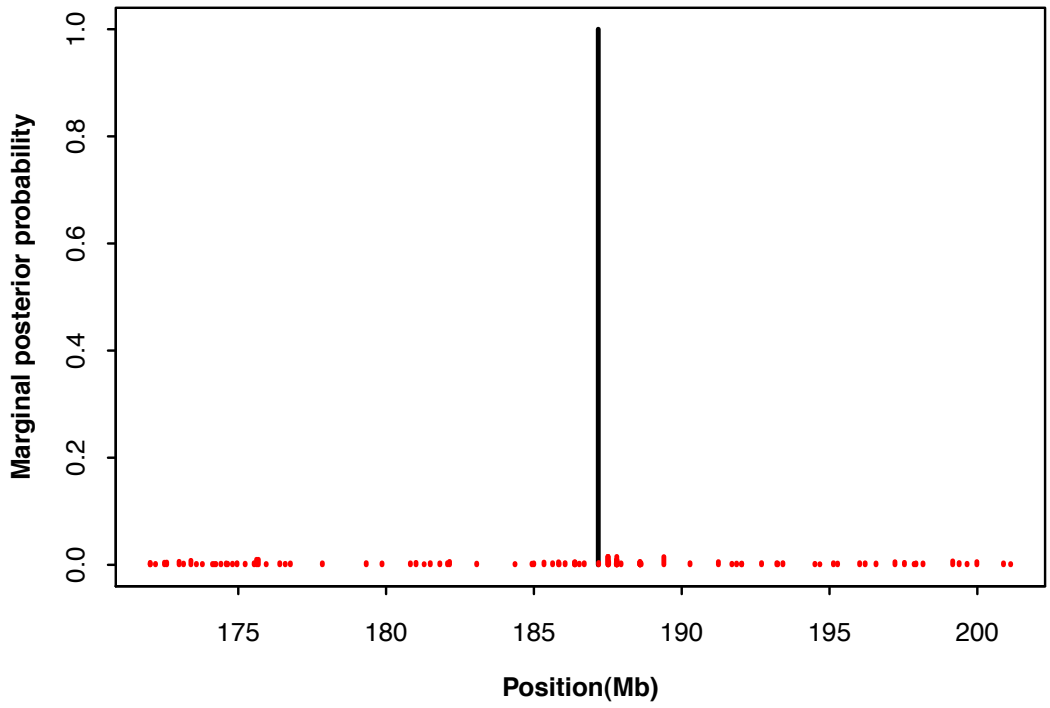


Figure 2.3 – The posterior probability of Bayesian association analysis along rat chromosome 1 using SNP data collected for population NW.

We evaluated a model statistically that posits that in the region for which we collected SNP data there is only on causal SNP underlying the resistance trait. To accomplish this we computed the posterior probability of association (PPA) for each SNP with the resistance trait and summarized the marginal PPA for models that describe association between the trait and one SNP only (1-SNP) and two SNPs (2-SNP). Our assumption was that it is unlikely that numerous, i.e. more than 2 SNPs, are located in this region under study. As shown in Figure 2.3, the SNP underlyignY139C located in *Vkorc1* has the highest marginal PPA of 0.98 and thus, is supported as the only causal SNP associated with the phenotype in the 30 Mb region studied.

Another classical feature of selective sweeps is that these cause an excess of linkage disequilibrium (LD) spanning a longer than expected chromosomal region, or LD blocks (Kim and Nielsen 2004). Here for the population NW population we observed that a LD block with high r-square values (>= 0.4) is supported for rat chromosome 1 positions 185-189 Mb, including the region where *Vkorc1* maps (Figure 2.2A). In contrast, in the fully susceptible population LH we found no support for a big LD block (Figure 2.2B).

For example, a SNP located in the gene *Ppapdc1a* (Ppapdc1a_188589203) is in significant LD with *Vkorc1* (squared r = 0.5) even though the sites are separated by a distance of ~ 1.4 Mb; Ld over such a distance is not expected under neutrality in the

absence of selection as we determined by forward-time simulations. Notably, whether we computed LD for our sample of rats available from the population NW, or for the sample of rats simulated by permutation of samples from population NW, significant LD blocks were found to flank *Vkorc1*.

2.3.4. Estimation of the selection coefficient on *Vkorc1*

We took advantage of the fact that we had samples available from different time points spanning 3 years in 14 natural populations. Notably, we had samples available from populations where rats were trapped prior to warfarin rodenticide treatment were collected and thereafter. Thus, across these time-series samples we were able to jointly estimate the selection coefficient s on *Vkorc1* as well as the effective population size N_e. Based on maximum likelihood analysis N_e was estimated as ~1,000. Notable, for the remaining analyses described the effect of N_e on s was small, at least as long as N_e was between 1,000 and 10,000 (Table 2.3).

When we considered all rat populations sampled the averaged selection coefficient against rodenticide susceptible rats was estimated as 0.347. However, this would include treatments with second-generation anticoagulants such as bromadiolone and difenacoum thought to be more toxic than warfarin. If we restrict our analyses to warfarin treatment only, i.e. only consider samples collected prior and after warfarin treatment (which in the field studies preceded treatments with bromadiolone and difenacoum), we estimated an average selection coefficient of ~0.295. Thus for the subsequent simulations and analyses we used $s = 0.3$ in our fitness models.

The frequencies of the mutant allele in *Vkorc1* decreased from 0.75-0.82 to 0.71-0.75 in three populations (populations 10, 12, 17) during the sampling period. Thus, the estimated selection coefficients in these populations appeared to be negative (Table 2.3). However, under a balancing selection model on an overdominant mutation this pattern is expected, as we discuss below.

Table 2.3 – Estimation of selection coefficient *s* on the Y139C mutation in *Vkorc1*.

PopID	N	R%	Assoc (P-value)	*s* (N_e=10^3)	*s* (N_e=10^4)	Initial frequency	End frequency	Poisons
4f(LH)	13	0%	-	-	-	0	0	No
11(KB)	70	83%	2e-4	0.310	0.273	0.68	0.73	W
21(SP)	23	33%	0.273	-	-	0.08	0	W
23(TH)	56	10%	0.001	0.254	0.273	0.07	0.17	W
5	20	74%	0.004	0.254	0.255	0.39	0.44	W
13	42	64%	2e-5	0.254	0.273	0.13	0.5	W
14	19	44%	0.001	0.401	0.428	0.18	0.5	W
12	25	95%	0.012	-0.033	-0.033	0.77	0.75	W+D
10	42	95%	0.170	-0.266	-0.263	0.82	0.73	W+D
4	52	94%	0.008	0.254	0.267	0.55	0.73	W+D
24(WU)	73	93%	9e-5	0.187	0.150	0.68	0.75	W+D
19	56	65%	3e-8	0.310	0.341	0.36	0.58	W+D
17	53	86%	1e-6	-0.033	-0.033	0.75	0.71	W+B
20	27	62%	3e-4	0.794	0.798	0.35	0.75	W+B
6	24	96%	0.103	0.545	0.481	0.85	1	W+B+D
28	37	86%	0.028	0.254	0.273	0.52	0.67	W+B+D
Average	40	69%	-	**0.347**	**0.346**	0.46	0.55	-

PopID: population ID and name abbreviation for 5 populations as in (Kohn, Pelz, and Wayne 2000).
N: the sample size in each population named by the farm number.
R%: the proportion of resistant rats in each population (the warfarin resistant level)
Assoc(P-value): the significance P-values of the genotype-phenotype association (CMH) tests.
s: corrected selection coefficient estimated assuming $N_e = 10^3$ and 10^4. '-' indicates *s* could not be estimated.
The allele frequencies of the Y139C mutation in *Vkorc1* at the beginning and the end of sampling period are provided.
Poisons: rat populations were exposed to different rodenticides, including warfarin (W), brodifacoum (B), coumatetralyl (C) and difenacoum (D) (c.f. Chapter 6).

LH population was not included for calculating average values of populations from the resistant area.

2.3.5. Allele frequencies of *Vkorc1* and linked sites

The frequency of the beneficial Y139C allele is expected to increase rapidly after warfarin selection. This is what we observed for the Y139C mutation in natural populations (Table 2.3). However, fixation of this adaptive mutation was not observed despite of the strong selection pressure; only in pop6 did the allele frequency of the *Vkorc1* resistance mutation increase from 0.85 to fixation after exposure to warfarin and the other two more potent second-generation anticoagulants difenacoum and brodifacoum (Table 2.3).

In contrast, and in general, we observed that allele frequencies at the end of the sampling periods (after warfarin treatment) approached 0.70-0.75 in 7 populations where resistance was prevalent (R% > 80%), and ~ 0.5 in 4 populations with intermediate resistance levels (population 5, 13, 14 and 19), and < 0.2 in 2 populations (TH and SP) with low resistance levels (R% < 33%) (Table 2.3). Thus, casual inspection of these results of allele frequency measurements suggests that the resistance trait cannot reach fixation. Similarily, the frequency of the beneficial allele decreases from 0.77 to 0.75 in Pop12, from 0.82 to 0.73 in Pop10 and from 0.75 to 0.71 in Pop17 (Figure 2.4). This pattern further suggests that the resistance allele approach an equilibrium frequency of ~ 0.75 in wild rats, which reminds us of the allele frequency change dynamics under balancing selection (Hartl and Clark 2007). Finally, the average mutant allele frequency in the highly resistant population NW

samples was estimated as 0.64, again, suggesting that the majority of resistance alleles are found in the heterozygote state.

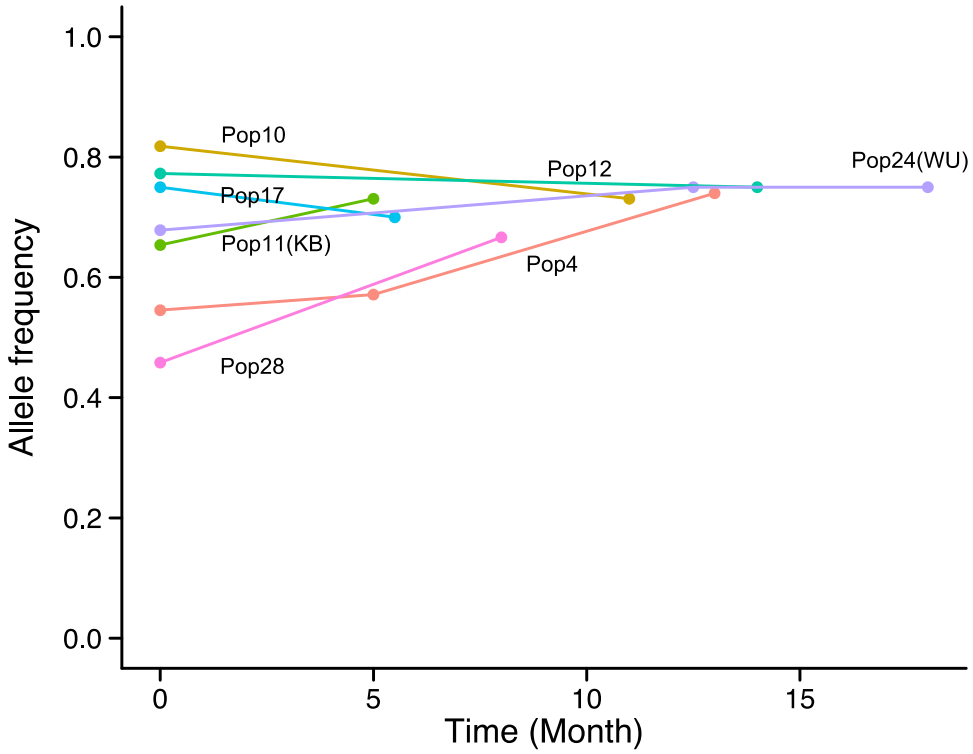


Figure 2.4 – Temporal allele frequency change of the *Vkorc1* Y139C variant in 7 natural populations of rats with resistance levels R% > 80%.

Our inferences of the haplotype frequencies at the onset of selection and their subsequent inferred change in frequency were consistent with the adaptive value of the Y139C bearing haplotype. Specifically, the allele frequencies of the neutral sites, which were initially linked to the beneficial mutation, increased. This was evident by comparing inferred haplotypes and neutral linked SNP allele frequencies collected

from the non-resistant population LH and the highly resistant population NW (Appendix 3).

2.3.6. Estimated recombination rates

We estimated the recombination rates in the regions flanking *Vkorc1* in the fully warfarin susceptible population LH and in the highly resistant population NW. Compared to a previous estimate of r = 0.006/Mb (Jensen-Seaman et al. 2004), the recombination rate estimated from LH population was found to be lower, ~0.002/mb. The former is used for the simulation of hypothetical population to show a general trend of genetic variation pattern, and the latter from LH is plugged in the empirical simulation for the comparison with the observation from empirical data (c.f. forward-time simulation section). The recombination rate based on NW samples is even lower, about 0.0006/Mb, which is expected after selection and might also be due to the potential haplotype structure in this population.

2.3.7. Forward time simulations

To assess whether the abundance change of the beneficial allele and the linked neutral alleles are due to genetic drift or are driven by selection, we performed forward-time simulation under neutral model and selection models. As shown in Table 2.1, without selection, the fitness values of genotype S/S, S/R and R/R are equally 1. Here, S represents the wild type nonresistant allele, and R represents the Y139C mutation in *Vkorc1* gene. If under warfarin selection, the non-resistant

genotype S/S has lower fitness ($1 – s$) and the selection coefficient s, as previously estimated, is ~ 0.3.

Moreover, we conducted simulations under two selection models to evaluate whether the beneficial mutation is not yet fixed under directional selection because it should not be fixed in this time scale since warfarin selection was introduced in the 1950s, or whether the lack of fixation is better explained by a balancing selection mode (Table 2.1). Under directional selection, we considered the dominance model with respect to warfarin resistance, in which both S/R and R/R genotype have fitnesses of 1 and the non-resistant genotype S/S has a fitness of $1 - s$ (~0.7). Under balancing selection on an over-dominant mutation the heterozygote genotype S/R has the highest fitness 1, S/S has the fitness of $1 - s \sim 0.7$, and the homozygote mutant genotype R/R has lower fitness than S/R due to the fitness costs (Greaves et al. 1977; Smith, Townsend, and Smith 1991; Kohn, Price, and Pelz 2008), represented by $1 - t$. Here we estimated t by the equation $P_R = s / (s + t)$. With the equilibrium allele frequency of the beneficial mutation $P_R \sim 0.75$ (Figure 2.4) and $s = 0.3$, t is estimated here as 0.1.

Here we further consider the question whether warfarin resistance has evolved by selection on a de novo mutation to Y139C in our study area or has evolved by selection on a standing variant that pre-dated the introduction of warfarin selection. , We simulated two scenarios to distinguish the two models.

To model the case of selection on standing variation we considered a scenario where the mutation leading to Y139C occurred on one allele ~700 generations ago;

i.e. before selection with warfarin selection but roughly corresponding to the time when *R.norvegicus* colonized Europe ~1716-1786 (Pelz et al. 2005). After evolving neutrally from generation -700 to generation zero selection with warfarin was added to the model. Generation zero thus was assumed to coincide with the 1950s when warfarin was introduced, which corresponds to about 100-200 generations (Adams and Boice 1983). We performed the simulations based on assumed allele frequency data across the 30 Mb region with spatially equally distributed SNPs each with the same initial minor and major allele frequencies. This simulation was done to explore differences between selective sweeps under the competing models.

To more quantitatively approximate expected values for allele frequencies and the chromosomal extent over which the selective sweep region can be seen we performed simulations that are based on the observed allele frequencies and actual chromosomal map locations of 100 SNPs linked to *Vkorc1*, again, with the initial allele frequencies obtained from the non-resistant population LH 150 generations ago (the generation when warfarin selection was introduced first). In sum, in both simulations we considered the alternative models of selection on new mutations versus standing variants, as well as the competing models describing directional selection versus balancing selection.

2.3.7.1. Balancing selection on an overdominant mutation Y139C

Patterns of allele frequency changes of the adaptive Y139C allele were examined by plotting the allele frequency change over time for the beneficial allele

under dominance and over-dominance models assuming the mutation arose as a new mutation or from selection on a standing variant (Figure 2.5).

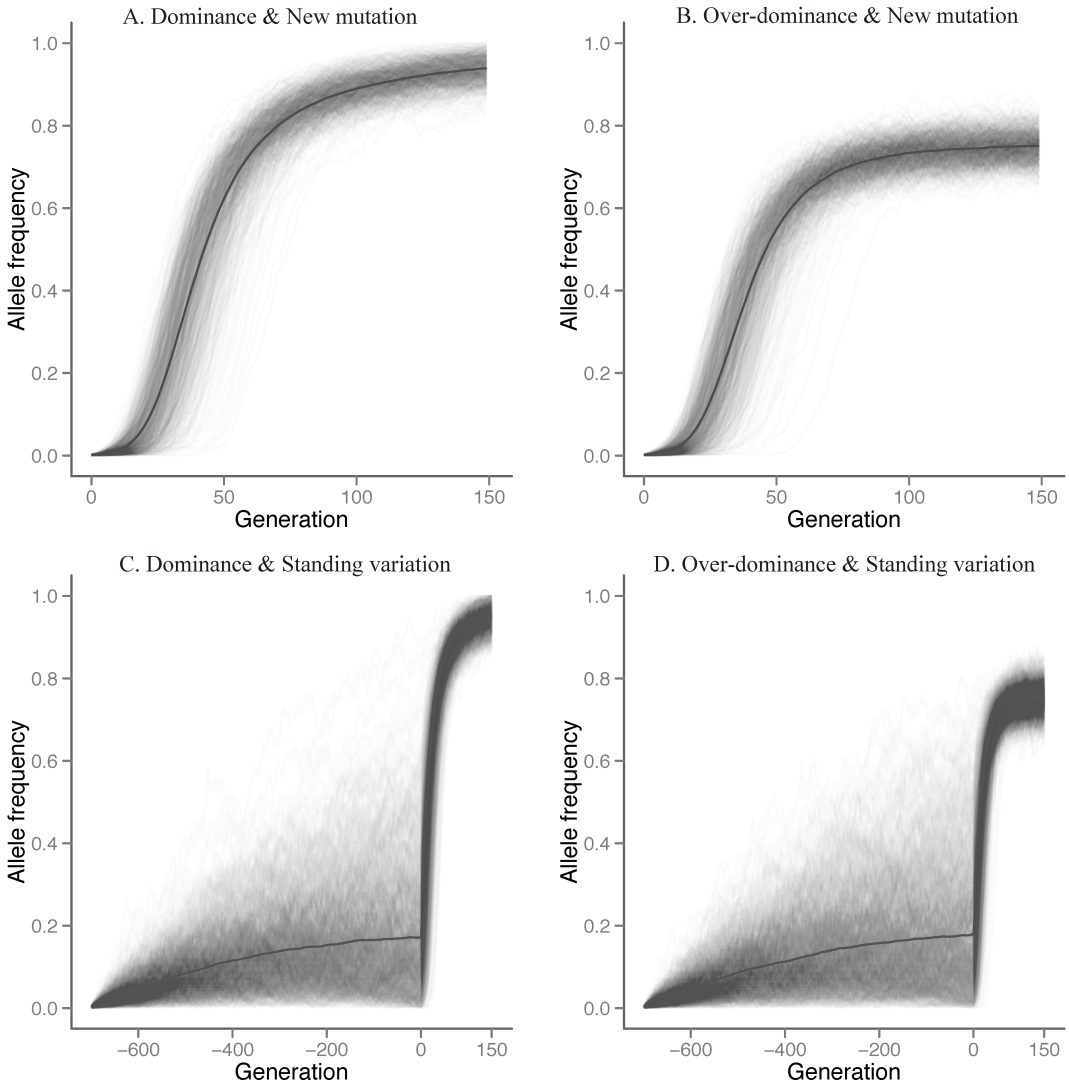


Figure 2.5 – Simulated allele frequency change of the Y139C mutation in *Vkorc1* under four selection models. (A) Dominance and new mutation model; (B) Over-dominance and new mutation model; (C) Dominance and standing variation model; (D) Over-dominance and standing variation model.

Patterns of allele frequency changes of the adaptive Y139C allele were examined by plotting the allele frequency change over time for the beneficial allele under dominance and over-dominance models assuming the mutation arose as a new mutation or from selection on a standing variant (Figure 2.5). Assuming selection on a new mutation under directional selection we observed that the Y139C allele frequency has already reached 0.9 under strong selection (e.g. selection coefficient = 0.3) as early as 100 generations (~ 25 to 33 years) after the introduction of warfarin selection (Figure 2.5A).

Under balancing selection, the frequency stays at the equilibrium frequency of ~0.75 after 100 generations (Figure 2.5B). Empirically observed Y139C allele frequencies were ~0.64 in the NW population and ~0.7-0.75 in most natural populations at then end of our sampling period (~150 – 200 generations after introduction of warfarin selection) (Figure 2.4). Thus, simulations of balancing selection more closely approximate the observed data. Notably, whetehr we simulated balancing selection on a new mutation or on a standing variant we would draw the same conclusion (Figure 2.5C and D). In addition, we demonstrated that the observed Y139C allele frequency could hardly be obtained under a neutral model (Figure 2.6).

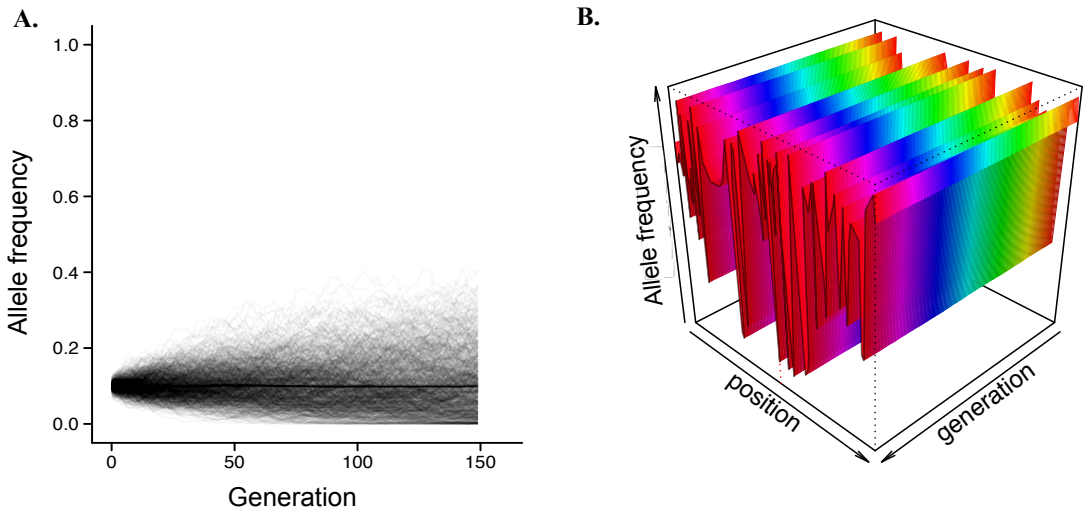


Figure 2.6 – Simulated allele frequency change under a neutral model (A – Y139C mutation and B – flanking region).

In the above analyses we considered the allele frequency change at Y139C. In the following we examine the allele frequency changes at SNPs that map in the flanking region of *Vkorc1*. In Figure 2.7 we plot the allele frequency change pattern across a 30 Mb region over 200 generations under warfarin selection. For the purpose of learning the different patterns among 4 models (see above), we performed the simulations with initial allele frequencies set to be 0.5 at each SNP. Simulations of directional selection on a new mutation revealed that the alleles linked to the beneficial allele have been driven to near fixation after 200 generations. In contrast, simulations of balancing selection on a new mutation resulted in an approach towards equilibrium frequencies for linked alleles. Directional selection can be distinguished from balancing selection as the former left a stronger signal than the latter. Specifically, under directional selection linked neutral alleles more rapidly increased in frequencies and the selective sweep region is longer (Figure 2.7A and B).

To examine results of simulations of selection on standing variants we plotted the allele frequency changes since the adaptive allele arose at -700 generations, and since it has been selected starting at generation 0. The comparison of results of simulations of directional selection with results obtained from simulations of balancing selection on a standing variant revealed also that under the former model allele frequencies more rapidly rise to high frequencies and the size of the selective sweep window is bigger (Figure 2.7C and D).

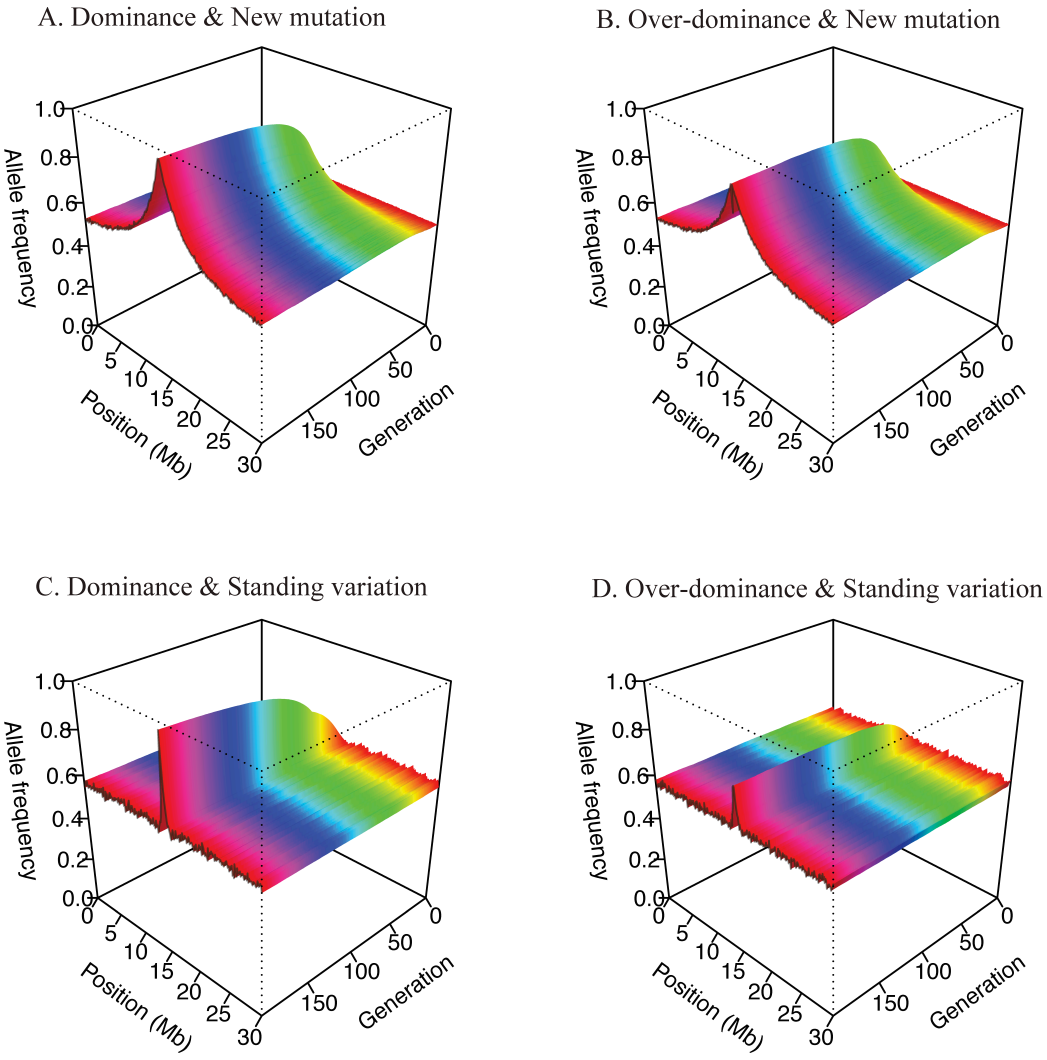
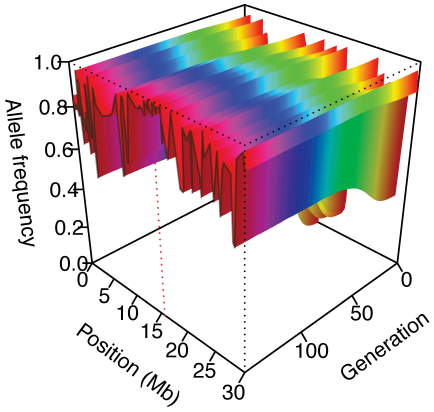


Figure 2.7 – Hypothetical simulation of allele frequency change of a sweep region across 30 Mb. A-D as in Figure 2.5 legend.

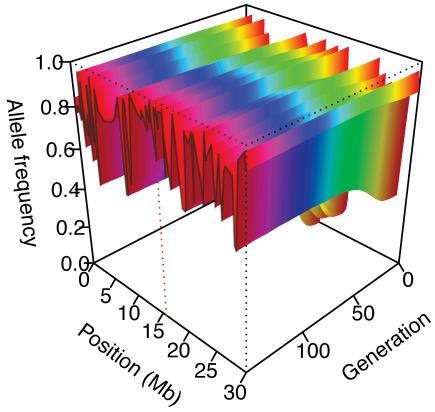
With the initial allele frequencies set as in the control population LH, the empirical simulation told a similar story as did the hypothetical simulations, where we started the simulations based on assumed allele frequencies as opposed to empirically observed allele frequencies in the susceptible population LH (Figure 2.8).

Specifically, we observed that both at the selected Y139C sites directional selection had a stronger impact on the linked neutral alleles also. We separately plotted the simulated allele frequencies for the final generation across the 30 Mb segment on chromosome 1 against the observed allele frequencies from population NW. The observations (gray wall) are most similar to the simulation results under the balancing selection model (Figure 2.8E). Besides, we showed that the observed pattern cannot be obtained under the neutral model, in which the allele frequencies of the whole region would merely be under drift (Figure 2.6B).

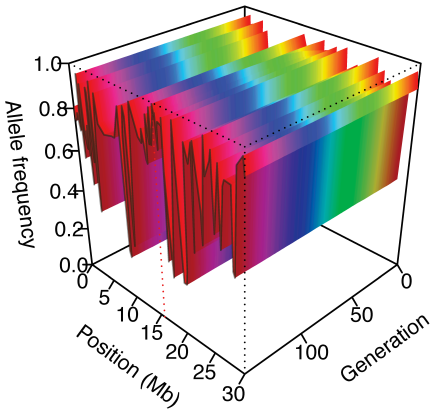
A. Dominance & New mutation



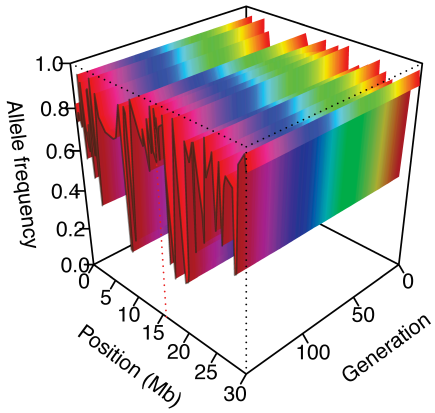
B. Over-dominance & New mutation



C. Dominance & Standing variation



D. Over-dominance & Standing variation



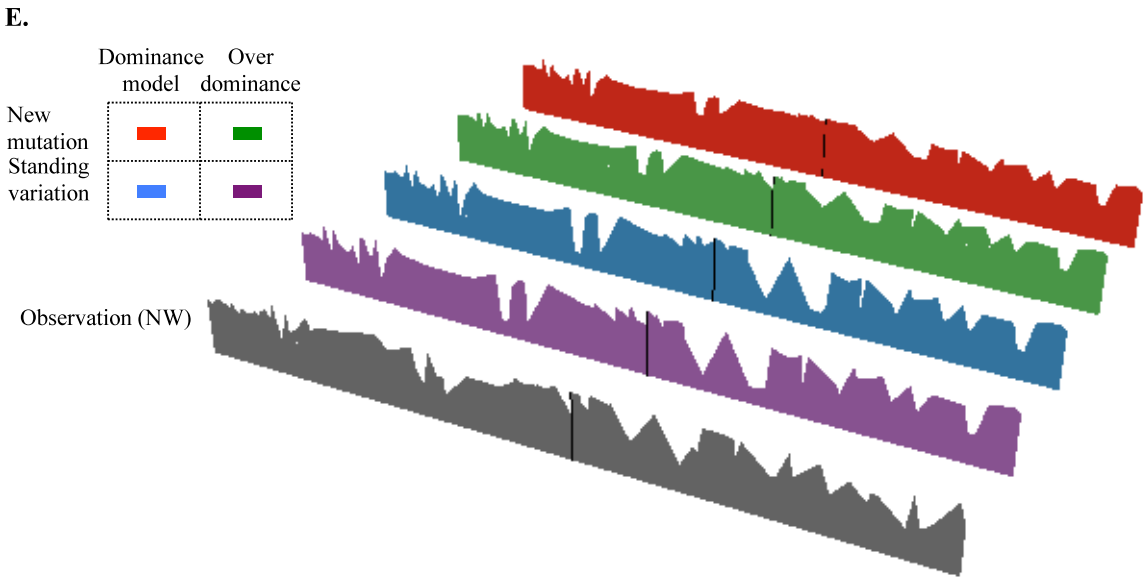


Figure 2.8 – Empirical simulation of allele frequency change. A-D as in Figure 2.5 legend. (E) The allele frequencies of the last generation across 30 Mb along the chromosome 1 under four models compared to observed data. The middle black vertical line indicates the position of *Vkorc1*.

2.3.7.2. Linked allele frequency dynamics support the Y139C mutation as a new mutation

It has been predicted that a selective sweep associated with selection on a standing variant is less pronounced in terms of reduction of heterozygosity and chromosomal distance over which this occurs than would be a selective sweep on a new mutation (Barrett and Schluter 2008). Our simulation results supported this theoretical prediction when we simulated these models under settings tailored to describe the evolution of warfarin resistance in rats in our study area (Figure 2.7 and Figure 2.8). When we simulated selection on a new mutation, we observed that most

SNPs in the 30 Mb region display strong hitchhiking dynamics. However, if we model selection on a standing variant that has already existed in the population for a long time before selection with warfarin (usually even greater than the -700 generations we simulated here), only a small region will display genetic hitchhiking dynamics. As expected, the precise outcomes depend on recombination rates assumed. Results of simulations confirm the expectation that selection on standing variants will affect more than one haplotype, and thus, even during the hitchhiking process recombination can occur, thereby resulting in more subtle drops of polymorphism and a smaller window affected by genetic hitchhiking.

An important finding during our simulations was that the initial frequency at SNPs linked to a pre-existing *Vkorc1* Y139C allele prior to selection matters. As an extreme case we a time prior to the use of warfarin and we consider an infinite population undergoing random mating. At this time the soon-to-be advantageous allele (Y139C) and flanking neutral alleles are in linkage equilibrium. Once the pre-existing beneficial allele will be selected the sweep signals in the flanking region would be strongly dependent on the minor and major allele frequencies at the linked neutral alleles, as these would be dragged along with probabilities that reflect initial frequencies, and these will not change after selection is introduced into the system (Figure 2.9C and D).

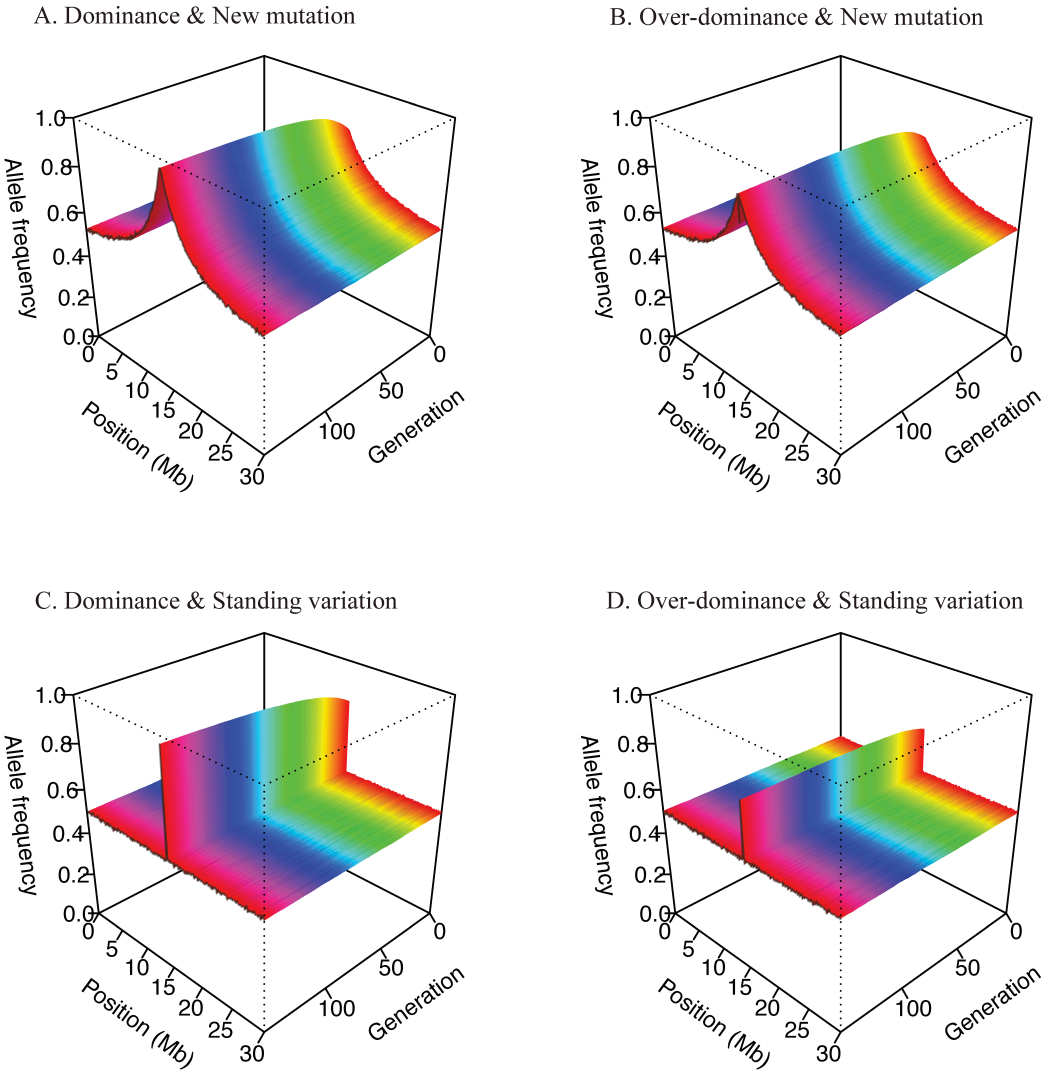
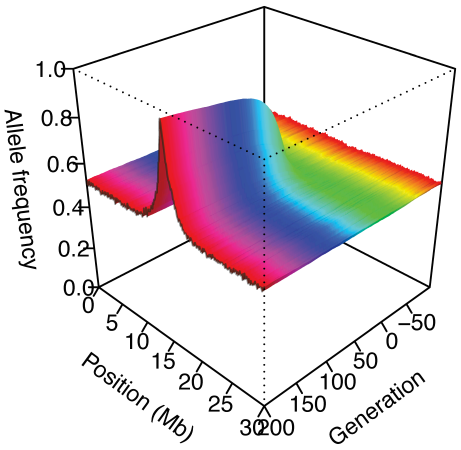


Figure 2.9 – Hypothetical simulation (under extreme scenarios without linkage disequilibrium before selection) of a sweep region across 30 Mb. A-D as in Figure 2.5 legend.

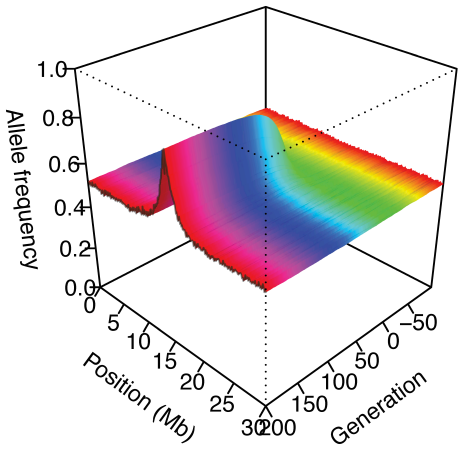
Thus, we expect that the longer the history of the Y139C adaptive allele has been prior to the introduction of warfarin selection the expected sweep effect on the flanking region once warfarin was introduced will be inversely related to the 'age' of

the Y139C mutation. To model this effect here we considered an age of Y139C as -700 generations and as -100 generations (selection starts at generation 0). We observed that this small difference in age is already discernible; the selection on the Y139C mutation of age -100 generations resulted in a stronger sweep signal than selection on the older standing variant of age -700 generations (Figure 2.10).

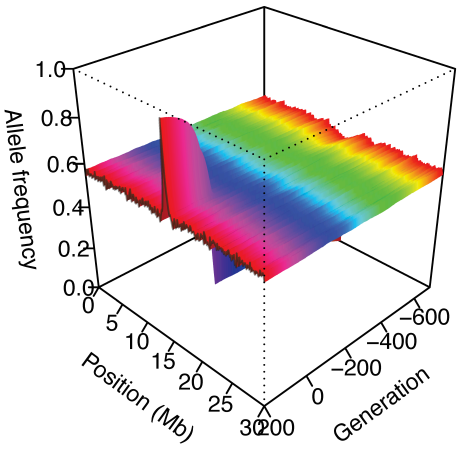
A. -100 generations (Dominance)



B. -100 generations (Over-dominance)



C. -700 generations (Dominance)



D. -700 generations (Over-dominance)

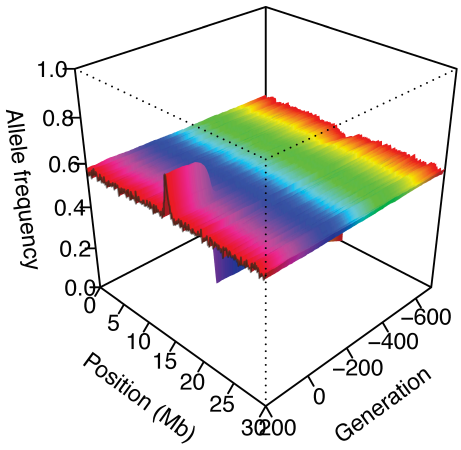


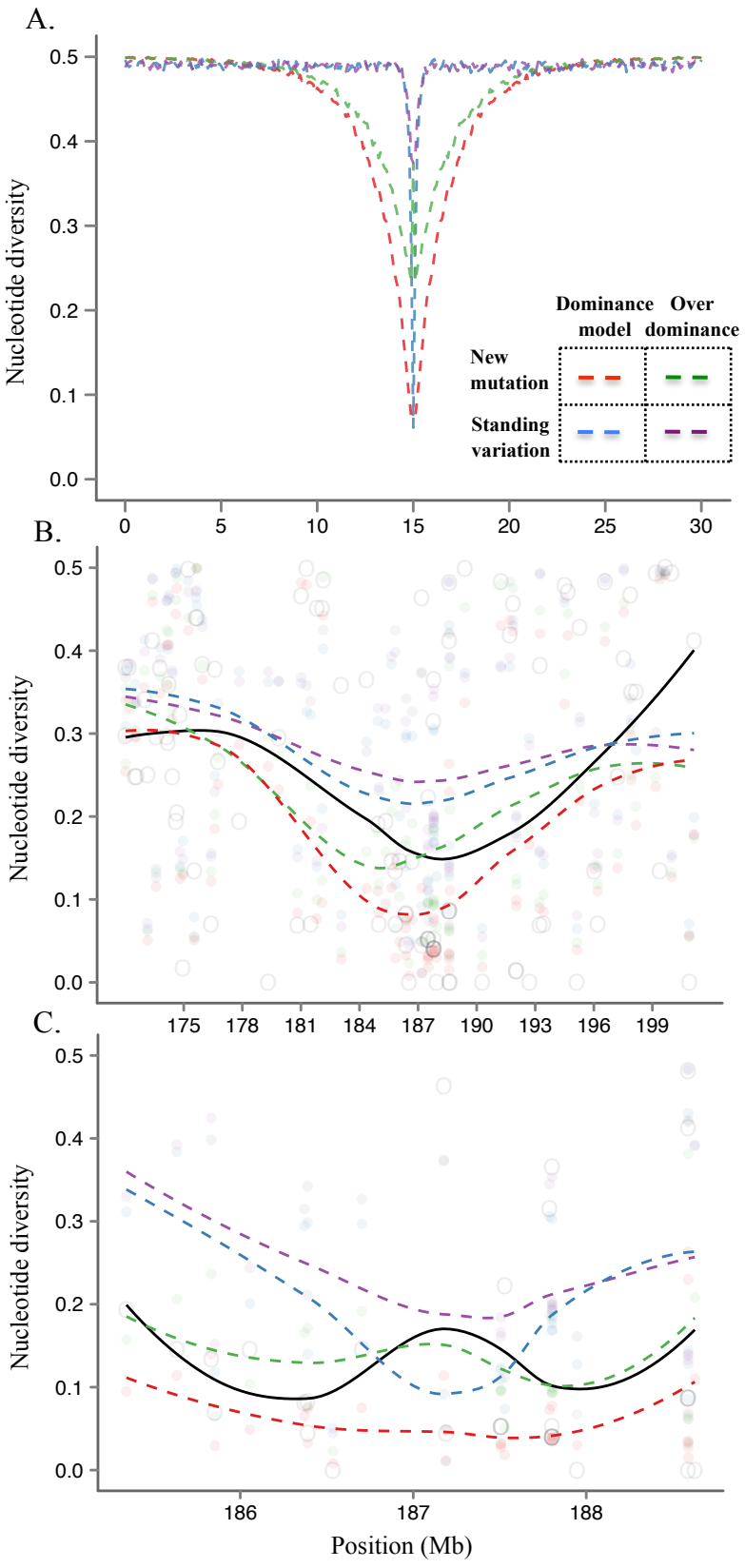
Figure 2.10 – Hypothetical simulation (considering -100 and -700 generations of the resistance mutation age before selection) of a sweep region across 30 Mb.

In Figure 2.8E, we plotted the observed SNP allele frequencies in the resistant population NW along the 30 Mb region as well as the SNP allele frequencies simulated under four selection models. The observations are most similar to the simulation results obtained under a model describing a new over-dominant mutation under balancing selection. Assuming a low recombination rate r = 0.002 during the simulations based on the observed allele frequencies in population LH genetic hitchhiking is only seen for SNPs that map within 3-4 Mb of the selected Y139C, both when we modeled balancing selection on a new variant and selection on a standing variant. But even for these SNPs nearby Y139C the relative increase in allele frequency was not enough to compare to the observed frequencies when we simulated selection on a standing Y139C variant. For example, the allele frequency of the linked SNP Ppapdc1a_188589203, which maps 1.4 Mb away from *Vkorc1*, increased from 0.09 in population LH to 0.71 in population NW that experienced warfarin selection (Appendix 3). This observation is much higher than the simulated frequency of ~0.55 from standing variation, but is in close agreement when simulating selection on a new variant, which reached ~0.70. Similarly, the window size affected by hitchhiking differed between when simulating selection on new and standing variants increase and the former results more closely resembled those observed (almost 30 Mb vs. ~ 2-4 Mb) (Figure 2.7 and Figure 2.8).

2.3.7.3. Reduced linked polymorphism supports Y139C as a new overdominant mutation under balancing selection

Selective sweeps are charccterized by local reductions of genetic polymorphism (Barrett and Schluter 2008). Both the size and shape of the sweep region contain information about the mode of selection and the age of the mutation. Here we compare simulated results and observed results for the polymorphism measures nucleotide diversity θ_π (equivalent to the expected heterozygosity) and observed heterozygosity H_{obs}.

As predicted by Figure 3 in (Barrett and Schluter 2008), under directional selection, the valley of reduced nucleotide diversity resulting from directional selection on standing variation is narrower than the valley resulting from selection on a new mutation (Figure 2.11A and B). We expanded these predictions by considering a balancing selection model also. We observed that the region of reduced polymorphism shaped by balancing selection on a standing variant is much shallower and narrower than those resulting from directional selection. The model describing directional selection on standing variation (blue line) has a narrower valley than the simulations of balancing selection on a new mutation (green line) when the entire 30 Mb-spanning region was considered (Figure 2.11A and B).


A.
Nucleotide diversity
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
30
Dominance model
Over dominance
New mutation
Standing variation
B.
Nucleotide diversity
0.0
0.1
0.2
0.3
0.4
0.5
175
178
181
184
187
190
193
196
199
C.
Nucleotide diversity
0.0
0.1
0.2
0.3
0.4
0.5
186
187
188
Position (Mb)

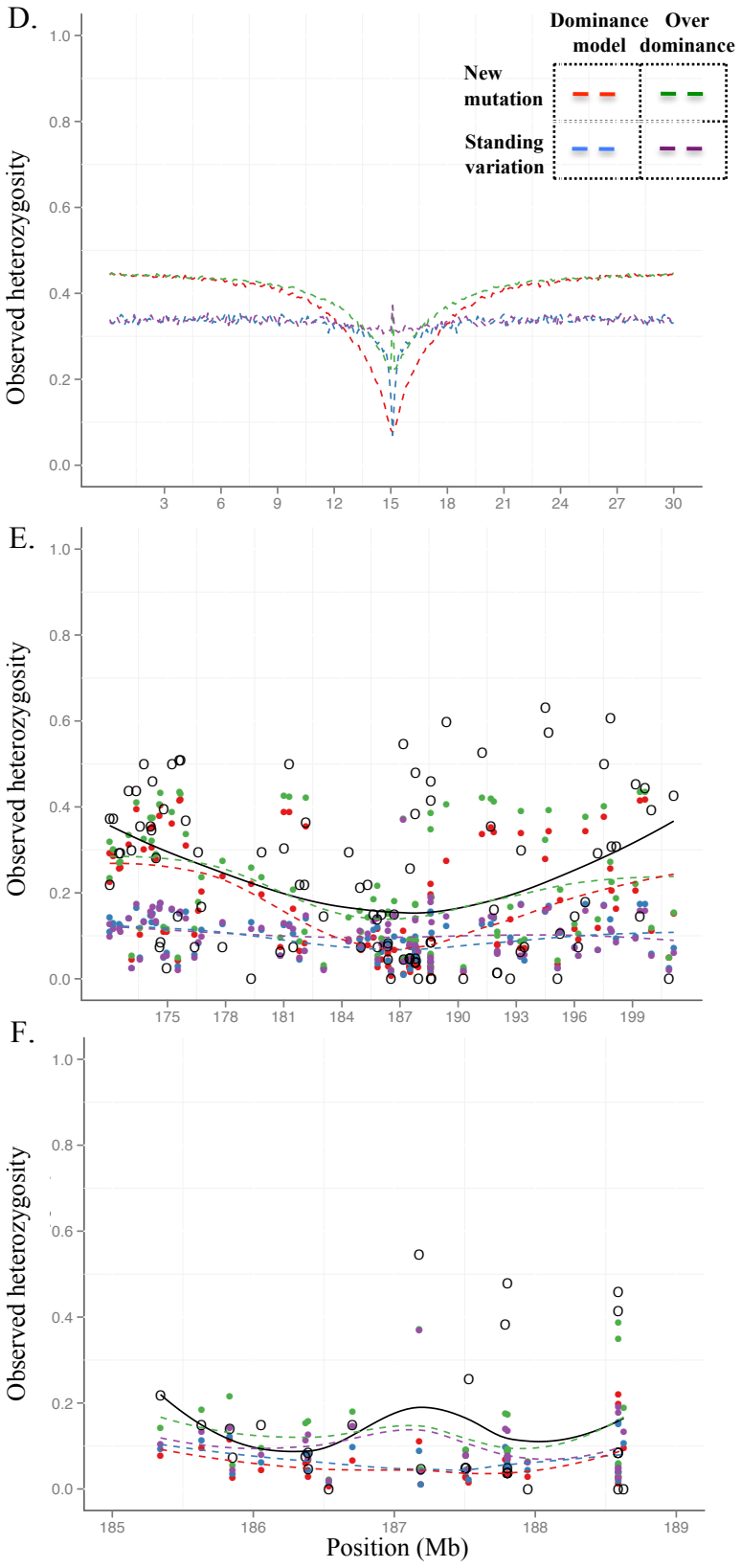

D.
Dominance model
Over dominance
New mutation
Standing variation
Observed heterozygosity
1.0
0.8
0.6
0.4
0.2
0.0
3
6
9
12
15
18
21
24
27
30
E.
Observed heterozygosity
175
178
181
184
187
190
193
196
199
F.
Observed heterozygosity
185
186
187
188
189
Position (Mb)

Figure 2.11 – Valley of reduced polymorphism. Nucleotide diversity of *Vkorc1* sweep region under (A) hypothetical simulation and empirical simulation: (B) 30 Mb sweep region and (C) the middle 4 Mb region. Observed heterozygosity: (D) hypothetical simulation and empirical simulation: (E) 30 Mb sweep region and (F) the middle 4 Mb region.

A more detailed analysis of 4 Mb region that directly flanks *Vkorc1* on both the 5' and 3' ends revealed even more pronounced such differences. Specifically, while the simulations of directional selection on a standing Y139C variant resulted in a valley of reduced linked polymorphisms, the simulations of balancing selection on a new Y139C mutation under balancing selection resulted in a valley of reduced polymorphisms also, but with a small peak of high polymorphisms at its center and where Y139C maps (Figure 2.11C; Table 2.1). Thus, the observed levels of nucleotide diversity (black line) are consistent with the simulated results describing balancing selection on a new mutation (Figure 2.11B and C).

The other measure of polymorphism H_{obs} supported the above conclusions (Figure 2.11D-F).

2.3.7.4. Patterns of linkage disequilibrium support Y139C as a new mutation

As described in Figure 2.2A and Figure 2.2B, an excess of linkage disequilibrium was observed surrounding *Vkorc1* in NW population, which suggested a recent footprint left by selection. As an example, we plotted the change of LD

between *Vkorc1* and the SNP Ppapdc1a_188589203 over generations under different models (Figure 2.12). Usually, the LD measure of D is the statistical covariance of haplotype genotypes, and r-square represents the square of statistical correlation coefficient (LaFramboise 2009). The best agreement with the observed r-square value 0.5 is the new mutation under over-dominance model, which supported our conclusion.

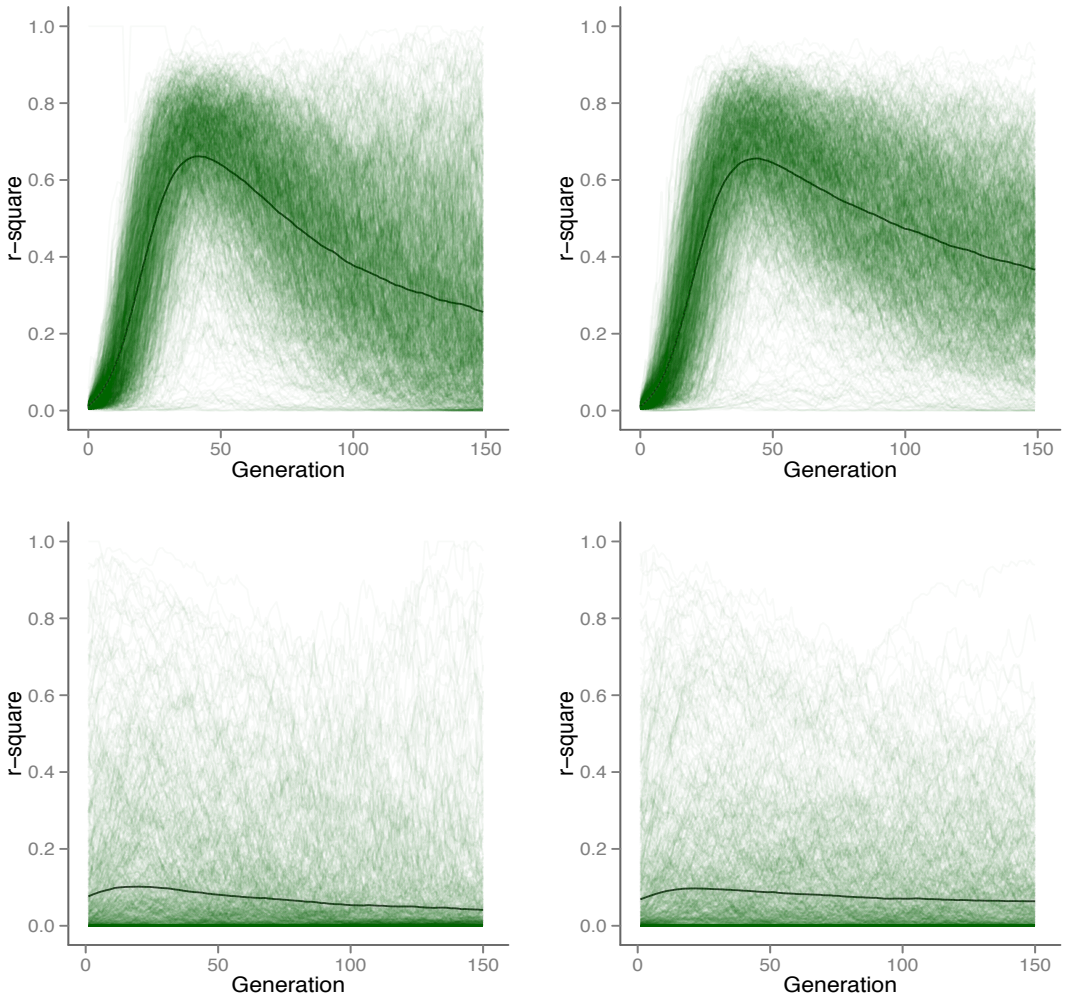


Figure 2.12 – simulated LD (between the *Vkorc1* and a neutral SNP) change over time (empirical simulation). A-D as in Figure 2.5 legend.

And assuming this model, we could roughly estimate the mutation occurrence time. Given the observed r-square value 0.5, the mutation would be occurred about 100 generations ago (Figure 2.12). From the end of our sampling year 1998, this estimation corresponded to 1965 – 1973, which is consistent with the early reported resistance rats in Germany at 1967 (population and agriculture 1986).

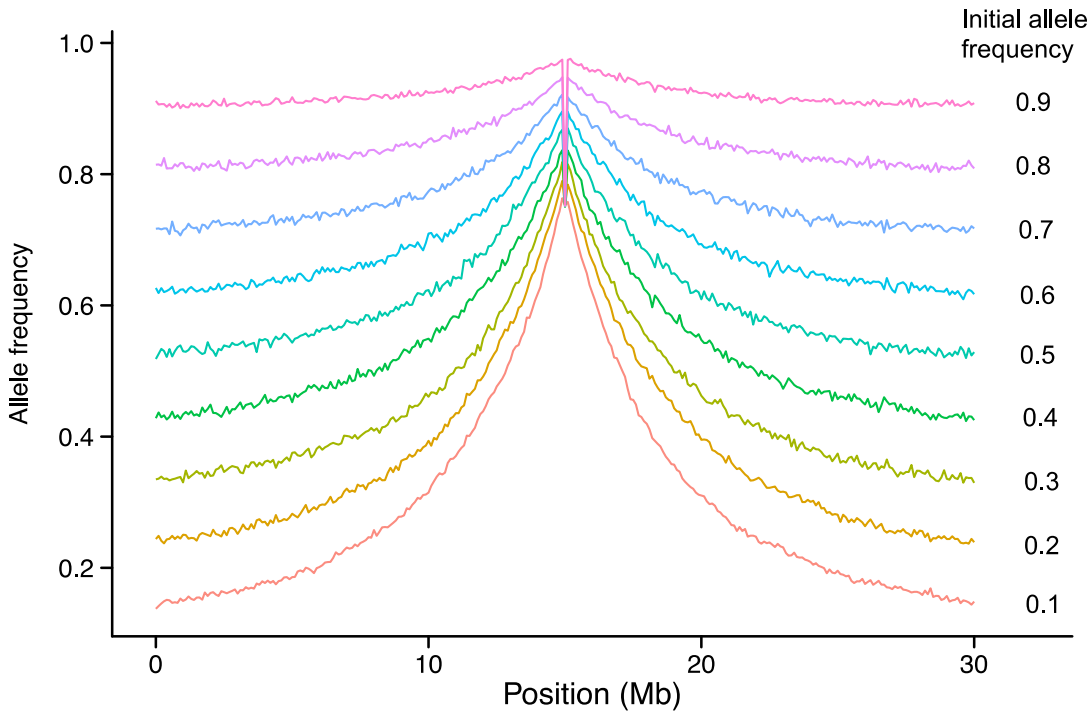
2.3.7.5. The effects of initial allele frequencies on sweep pattern?

Neutral SNPs with different initial allele frequencies will respond differently to the "dragging strength" of selection. This is a somewhat surprising finding in this work and hasn't been paid enough attention before. Sweep seems to have larger effects on neutral alleles with low frequency than high frequency. Although in general, the beneficial mutation has higher chance to occur on the chromosome that the neutral alleles have high frequencies. And this is what we observed after inferring the neutral alleles linked to the adaptive mutation in *Vkorc1*. Examining the initial allele frequencies at the SNPs that are linked to the beneficial allele we observed that only 17% of these have allele frequencies < 0.5 in population LH. In contrast, in NW population most SNPs that are linked to the adaptive mutation have high frequencies; with 64% of these displaying major allele frequencies of >= 0.7.

Here, we observed that neutral SNPs with an initially low frequency tend to increase rapidly by genetic hitchhiking. For example, the SNP Ppapdc1a_188589203 has the initial frequency of 0.09 in the susceptible population LH, but increased to 0.71 in the highly resistant population NW. In contrast, its neighboring SNP Ppapdc1a_188589271 with an initial frequency of 0.23 only increased to 0.52 after

selection. Such differences in the initial frequencies also affected LD. The r-square value between the low frequency SNP Ppapdc1a_188589203 and *Vkorc1* is 0.5, which is much higher than the r-square between its neighboring high frequency SNP Ppapdc1a_188589271and *Vkorc1* (r-square = 0.25). The general patterns followed these examples. Specifically, we examined the allele frequencies at SNPS at the last generation with regard to the different initial frequencies chosen for the simulations (0.1-0.9), while modeling balancing selection on a new overdominant Y139C mutation, by plotting the allele frequency itself (Figure 2.13A) and the allele frequency changes (delta allele frequency; which we calculated as observed allele frequency minus the initial frequency, Figure 2.13B). We observed that linked neutral allele frequencies increase much more rapidly for alleles with low initial frequencies.

A.



B.

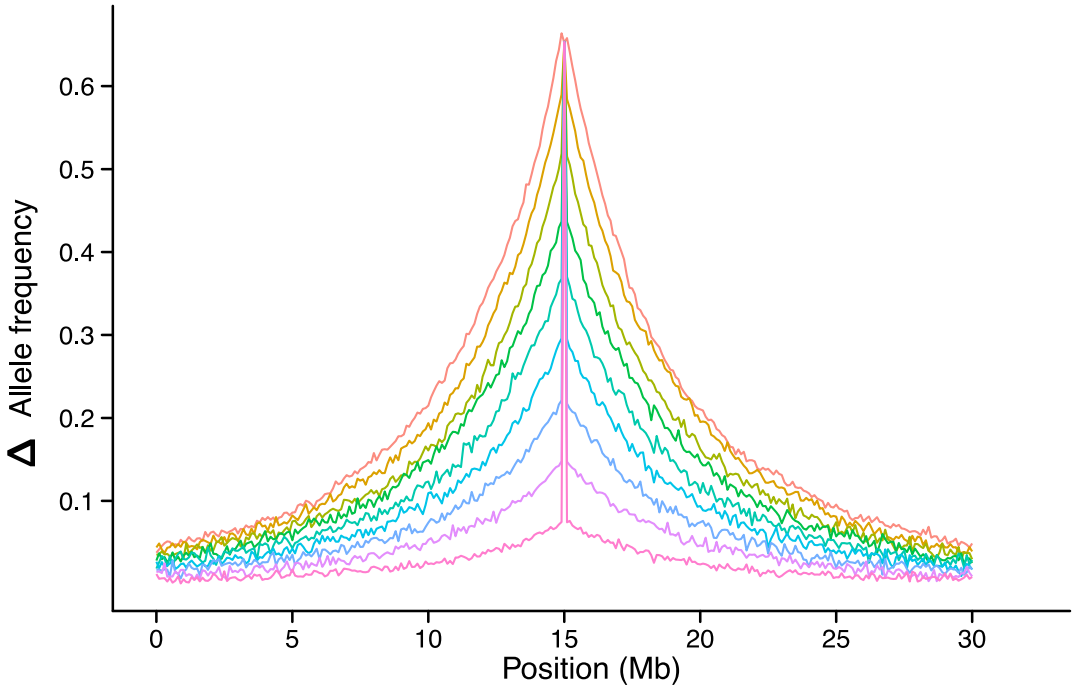


Figure 2.13 – The effect of different initial allele frequencies on selective sweep pattern (hypothetical simulation). (A) Allele frequency after selection and (B) delta allele frequency: allele frequency after selection minus before selection.

Different settings of the intitial SNP frequencies affected profoundly affected expected patterns of LD. We plotted r-square between SNPs linked to Y139C across the 30 Mb region from generation 0 to generation 150, with the initial allele frequencies set as observed in the susceptible population LH (empirical simulation, Figure 2.14 and Figure 2.15). Generally speaking, LD decreased with distance from the selected locus. However, closely linked SNPs with different initial allele frequencies exhibited big difference in LD. Specifically, we observed several small LD peaks amongst what otherwise resembled a plateau of low to intermediate LD. We attribute these to those few SNPs where the initial frequency of the alleles linked

on the same haplotype of Y139C was low.

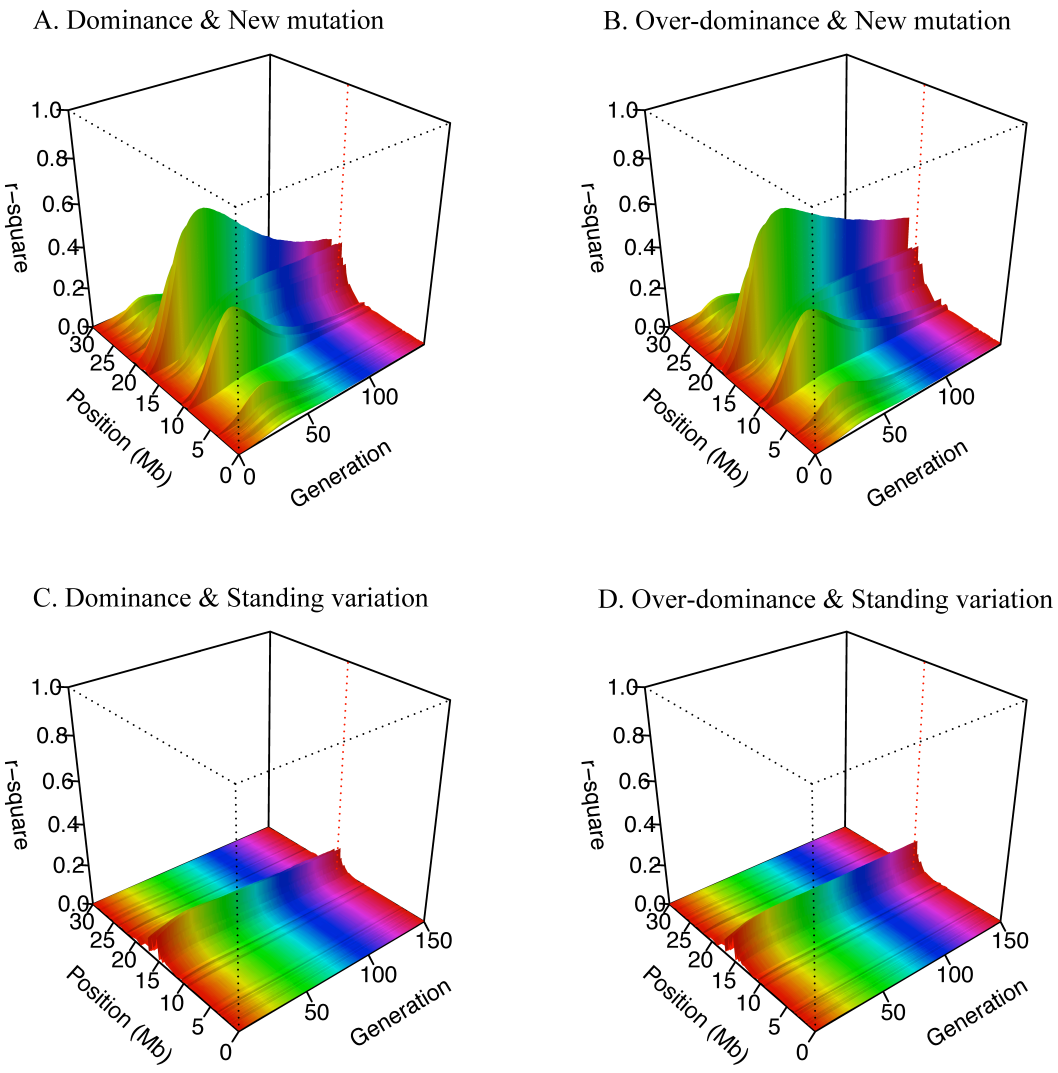


Figure 2.14 – LD (between the adaptive variant and surrounding sites) change over time across 30 Mb sweep region (empirical simulation, backside view). A-D as in Figure 2.5 legend.

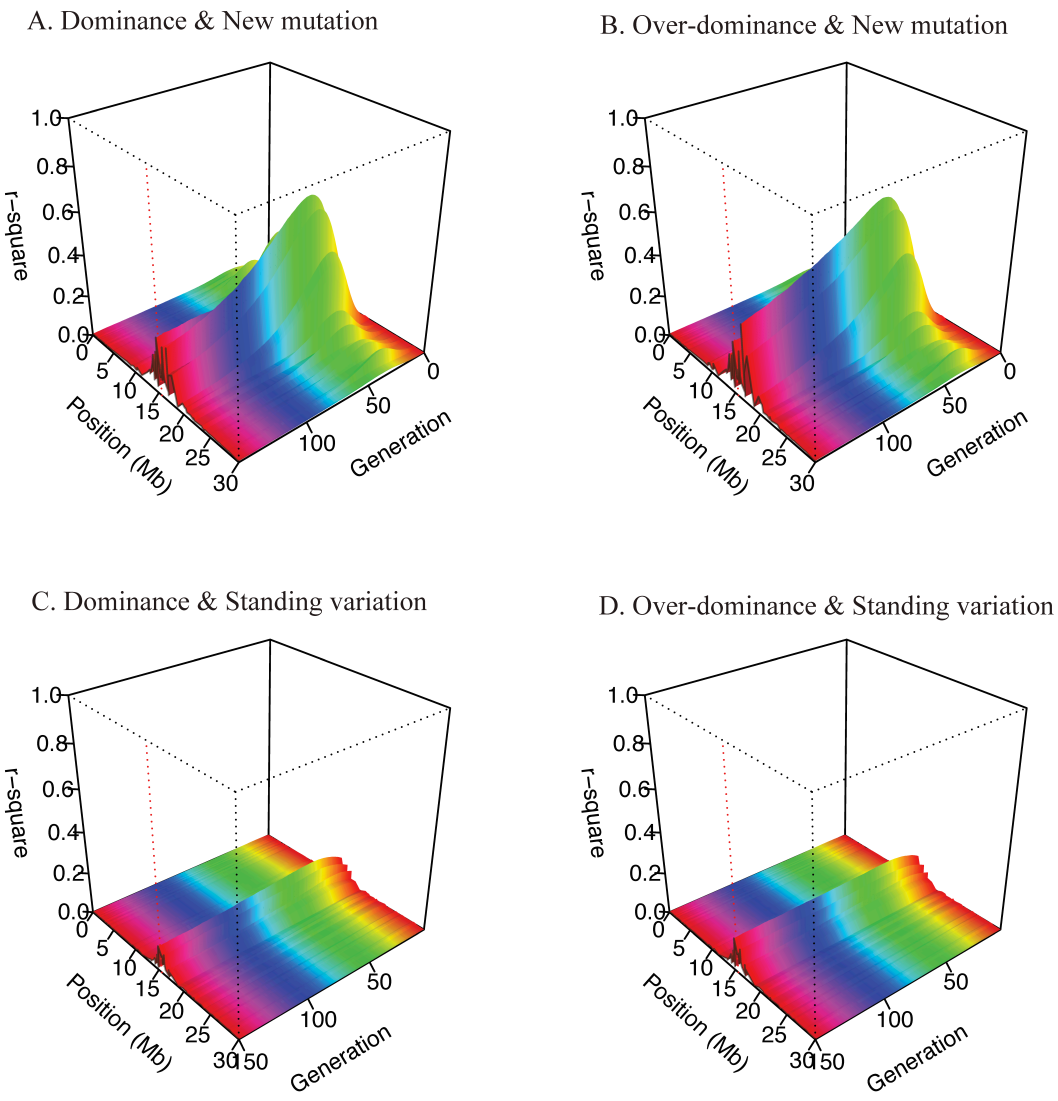


Figure 2.15 – LD (between the adaptive variant and surrounding sites) change over time across 30 Mb sweep region (empirical simulation, frontside view). A-D as in Figure 2.5 legend.

2.3.7.6. How long the sweep spans?

Consistent with previous studies (Kohn, Pelz, and Wayne 2000) our simulations revealed a region on chromosome 1 that is affected by the selective

sweep associated with warfarin selection on Y139C that spans ~30 Mb when we modeled balancing selection on an overdominant locus with $s = 0.3$ (Figure 2.8, Figure 2.11, Figure 2.15). To generally estimate chromosomal region affected by a selective sweep under various selection intensities, we conducted the simulations based upon assumed (set) prior allele frequencies with a recombination rate of 0.006 Mb estimated from the laboratory rat strain. We simulated the models describing selection on a new mutation under directional selection with s chosen to vary between 0.01 and 0.50. Balancing selection affects a narrower region as indicated in Figure 2.7. Here we only show the directional selection case for future reference purpose, as it is the commonly assumed model in most genetic analysis.

As expected based on theory, weaker selection strengths require longer times to drive the new adaptive mutation to high frequency. However, it is indicated that for $s >= 0.3$, 200 generations are enough for increasing to increase an adaptive mutation almost to fixation; if we assume $s = 0.1$, it requires 400 generations. For $s = 0.05$ and 0.01, we found that 1,000 and 2,000 generations are needed for an adaptive mutation to reach fixation, respectively.

Clearly the assumed selection coefficient is of importance in any system under selection. Here we plotted the allele frequencies across the whole 30 Mb region assuming a hypothetical adaptive allele has reached fixation under different selection pressures ($s = 0.01$, 0.05, 0.1, 0.2, 0.3 and 0.5). As indicated in Figure 2.16, when $s = 0.01$, only a small region (within 2 Mb) will be affected; whereas when $s = 0.5$, the

sweep region is even greater than 30 Mb. This prediction with varying selection coefficients would be considered as a reference in future search for sweep regions.

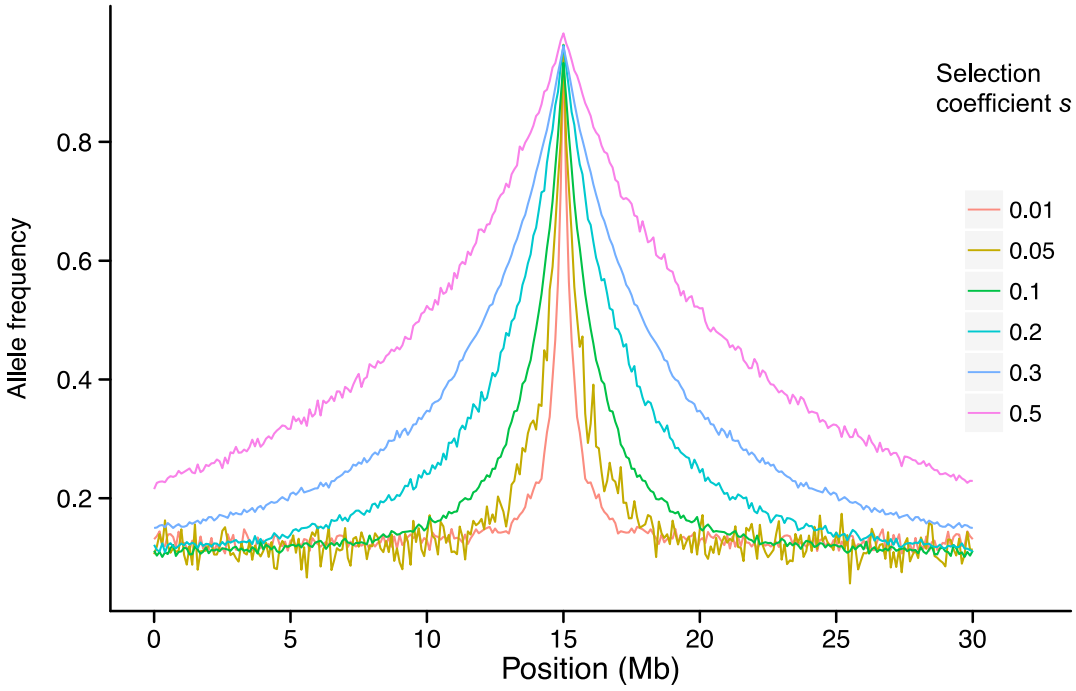


Figure 2.16 – The effect of different selection strength on selective sweep pattern (hypothetical simulation). Simulation is performed under directional selection to provide reference for further studies as directional selection is the most commonly assumed model.

2.3.7.7. Association between *Vkorc1* and warfarin resistance cannot be obtained randomly based on simulation

The penetrance of the Y139C mutation in *Vkorc1* has been estimated based on the proportions of resistant rats with each of the resistance genotypes from an analysis of the resistant population NW and the other natural populations available for study

(Table 2.1). About 93% rats with the SS genotype were susceptible to warfarin. In contrasts, >95% rats with the SR or RR genotypes were warfarin resistant. This near complete penetrance with regard to warfarin is used in our simulations and corresponds to previous such estimates of penetrance (Kohn, Pelz, and Wayne 2000). Notably, the penetrance of *Vkorc1* Y139C with regard to bromadiolone and difenacoum poison resistance is much lower and closely resembles co-dominance (Kohn, Pelz, and Wayne 2000).

Our analysis of the observed data showed that the Y139C *Vkorc1* SNP under balancing selection is significantly associated with the resistance phenotype (P-value in 1,000 replicates $< 4\times10^{-5}$). We could not replicate this result when we simulated the neutral process (P-value in 1,000 replicates = 0.05). Even when we simulated directional selection on Y139C we could not re-capitulate this level of significance of association between Y139C and the warfarin resistance trait as measured using the BCR method. In fact, since under a directional selection model the beneficial allele is almost fixed the association tests returned non-significant P-value (1,000 replicates) of > 0.1, as the observed resistance frequency (and considering penetrance) was much below fixation (~0.7).

2.4. Discussion

"Evolutionary adaptation is a special and onerous concept that should not be used unnecessarily, and an effect should not be called a function unless it is clearly produced by design and not by chance."

George C. Williams, 1966.

As pointed out by Barrett et al. in 2011, great progress has been made in identifying trait-related genes, but less is known about the details of the selection regime (Barrett and Hoekstra 2011). By examining the details of a specific selective sweep it is not only possible to identify the "adaptive" variations, but furthermore, to test for these generally unknown details of the selection regime.

Prominent examples that first documented to possibility to detect "adaptive" variations via genome scans for selective sweep signals included the early efforts to map the warfarin resistance locus (Kohn, Pelz, and Wayne 2000) and adaptive mutations in *Drosophila* (Harr, Kauer, and Schlötterer 2002). Other such examples have followed, and these include the identification of beneficial alleles associated with lactase tolerance in humans (Tishkoff et al. 2007) and genes that conferred insecticide resistance in *Drosophila* (*Cyp6g1*) (Kohn, Pelz, and Wayne 2000; Schlenke and Begun 2004; Tishkoff et al. 2007).

Despite these successes, the designation of "adaptive" in many cases is premature and the underlying selective mechanisms are unknown. Specifically, factors that shaped the selective sweep, such as the selection strength, the fitness model and mutation age (relative to environmental changes), generally have not been explicitly considered.

Here by testing the genotype-phenotype association of the known warfarin resistance gene *Vkorc1* and by investigating the genetic variation surrounding the

gene we connected genotype, phenotype and fitness; as it is required to fully document the role of an adaptive allele underlying an adaptive trait (Barrett and Hoekstra 2011). In particular we employed rigorous forward-time computer simulations (Kim and Stephan 2002; Peng, Amos, and Kimmel 2007; Servin and Stephens 2007; Carvajal-Rodriguez 2008; Hoban, Bertorelle, and Gaggiotti 2012) to distinguish among several scenarios that could have promoted the evolution of warfarin resitance in the Norway rat. Such forward-time simulations have critical advantages over backward (coalescent based) simulations in predicting outcomes, testing hypothesis and evaluating statistical approaches (Hudson 2002; Peng, Amos, and Kimmel 2007; Carvajal-Rodriguez 2008; Kim and Wiehe 2009).

In this study, we took advantage of the forward-time population simulation computational environment simuPOP (Peng and Kimmel 2005) to compare competing hypotheses that could explain the evolution of warfarin resistance as mediated by the Y139C mutation in *Vkorc1*. Although the Y139C mutation in *Vkorc1* cannot explain all the aspects of resistance, it has been identified as the primary resistance factor that promoted the spread of warfarin resistance in our study area in Germany (Pelz et al. 2005). The comparison between simulated results under different models and the empirical data supported a model positing balancing selection model and a new or young overdominant Y139C warfarin resistance mutation in Norway rat populations that we study in Germany.

2.4.1. Selection strength and fitness models

Measurements of selection intensities and relative fitness ratios done during field work in the 1970s established warfarin resistance in the Norway rat as a now classical example of balancing selection on an overdominant mutation (Greaves et al. 1977). In the age of genome scanning and population genomics, and with the knowledge of the warfarin resistance gene in hand these predictions can be tested at a more rigorous level. In addition, our work consists of field work over many years, with the sampling of rats from prior and after warfarin selection, and genome-scale SNP diversity data. Thus, we can combine field data and molecular data to estimate crucial parameters, unlike most other methods that infer selection coefficients based on the genetic diversity data around presumably adaptive sites of unknown adaptive significance (Kim and Stephan 2003; Barrett and Hoekstra 2011).

Specifically, with genetic polymorphism data in hand for rats that were sampled at multiple time points during a three-year field experiment, we were able to jointly estimate the selection coefficient and the population size using Bollback et al.'s maximum likelihood approach (Bollback, York, and Nielsen 2008).

Our estimation of the selection coefficient of 0.3 for the wild type homozygous rats under warfarin selection is similar to previous calculations of 0.32 (1-0.68, c.f. Table 2.1) (Greaves et al. 1977). This supports the value of filed work and the monitoring of phenotype frequencies in the field, but also lend credentials to molecular population genetic studies estimating *s* in the absence of any field data.

And the following simulation with selection coefficient of 0.3 in the fitness model is in concert with our observation. The selection coefficient of ~0.3 estimated for the warfarin resistant rat populations from our study area in Germany is high, above that estimated for most natural populations: for example, *s* estimated against the wild type genotype in human populations where malaria is common is ~0.1 (Hartl and Clark 2007) and *s* for alleles promoting lactase tolerance are ~0.04-0.097 (Barrett and Hoekstra 2011). The selection coefficients estimated for the two genes associated with coat color variation after horse domestication are about 0.0007 (*Agouti*) and 0.0019 (*Mc1r*) (Ludwig et al. 2009). Selection on coat color variants in beach mice are ~0.07-0.21 (Barrett and Hoekstra 2011). Thus, our study system is somewhat extreme in terms of the selection regime. However, our main aim to study this system in more depth is to establish a study system that can be used to more deeply investigate the genetics of adaptation. Notably, the study system was amongst the first to illustrate the power of population genomic screens to identify and localize adaptive variants in natural populations.

For the resistant heterozygotes and homozygotes, the observed allele frequency change patterns, and the simulations, supported a model that posits balancing selection on an overdominant variant. The fitness cost parameter *t* associated with the warfarin resistant homozygote was estimated as $t = 0.1$ using the 'known' selection coefficient of ~0.3 and the equilibrium frequency of the adaptive Y139C allele frequency of 0.75 (Table 2.1). This estimate is much smaller than the previous estimate of 0.63 (1-0.37, c.f. Table 2.1; (Greaves et al. 1977)). However, the

observed allele frequency at Y139C and the reduced levels of linked polymorphism agree well with the simulation results we obtained by adopting $t = 0.1$.

The fitness cost of the resistant mutation has been studied for years and it may involve multiple pathways. Accumulated researches reported high requirements of Vitamin K, blood clotting disorder and aorta mineralization in resistant rats (Markussen et al. 2003; Kohn, Price, and Pelz 2008). Recent experiment demonstrated that at vitamin K deficiency, blood clotting time increased in homozygous resistant rats, but only a little in heterozygous males, not in heterozygous females or wild-type rats (Jacob et al. 2012). This study specifically supported the over-dominance model since heterozygotes would have higher fitness in wild environment when Vitamin K is inadequate. Here, we not only estimated the relative fitness ratio for three genotypes of *Vkorc1* gene, we also showed the genetic evidence for the balancing selection model. Directional selection is the commonly assumed and tested model, but the heterozygote advantage selection is believed to be acting on a small portion of loci and short-lived (Hedrick 2012). As we are exploring the general role of balancing selection in shaping the genome, there is argument that over-dominant alleles might be more common than we previously thought and play important roles in adaptation (Cochran et al., in progress).

2.4.2. The effect of mutation age on sweep pattern

There is no absolute cut between new mutation and the standing variation. Instead, it could be viewed as a continuous changing trend. “Old” standing variation that has existed in the population with long history would have weak effect on

flanking regions after selection. The younger age of the variation, the stronger effect of sweep (Figure 2.10). And if it is a de novo mutation after selection, it will affect a widest region surrounding it. Here a new mutation in certain population could be either a new 'born' or a variation introduced by migration.

2.4.3. Linkage disequilibrium under balancing selection

In this first SNP based analysis, directional selection on Y139C caused pronounced LD blocks surrounding the selected site, as was expected based on theory (Kim and Nielsen 2004; Stephan, Song, and Langley 2006) and based on previous studies on the same rat populations using microsatellites (Kohn, Pelz, and Wayne 2000). Balancing selection, however, is not expected to cause this much LD (Charlesworth 2006). However, if the selection coefficient underlying balancing selection is high, such as it is in this study system, the simulations conducted here revealed that strong LD pattern can result from over-dominance (Figure 2.2A).

The age of the Y139C mutation is unknown, but the rapid evolution of resistance as mediated by this mutation raised speculations as to the age of this mutation; i.e. whether the mutation pre-dated the introduction of warfarin selection. Unless better sampling of rats from their ancestral areas on Asia is done, and museum samples are analyzed, this issue remains a subject of speculation. However, by comparing the observed levels of LD with the simulated levels of change of LD over generations it is possible to discern the age of the mutation to some degree. Specifically, by assuming a range of recombination rates and selection coefficients the LD change over time can be modeled using forward-time simulations (Figure

2.12). Following this approach our estimations were consistent with a selection model that posits balancing selection n an overdominant Y139C mutation that occurred, or arrived in form of a migrant, at about 1965-1973. This estimate matches up well with the documented first occurrence of resistance in Europe (population and agriculture 1986).

2.4.4. The initial allele frequencies of linked neutral alleles strongly affect the expected selective sweep

The selection intensity, time scale, and levels of recombination strongly affect the power to detect selective sweeps (Nielsen et al. 2005). In this study we considered numerous models and parameter settings to compare the expected signature of selective sweeps associated with warfarin resistance with those observed.

Amongst the numerous factors that affect the signature of selection on genes and their neighboring loci, of these, the initial frequency of alleles at linked sites here was shown to impact analyses and expected outcomes in a most profound fashion.

Similarly, the impact the ‘age’ of the adaptive allele has, and thus the ‘age’ of the neutral alleles linked to it, profoundly affected the outcomes of the simulated selection scenarios. We showed here for simulated data and for the observed data collected from a warfarin susceptible population (LH) that SNPs linked to ‘young’ low frequency alleles display drastic patterns of genetic hitchhiking; much unlike ‘old’ high frequency SNP alleles that have been on the same have as the wild type allele at the selected locus (Figure 2.13 and Figure 2.14).

2.4.5. Conclusion

In this study, we established the association between the Y139C mutation in *Vkorc1* and warfarin resistance. We further provided estimates deduced from the observed polymorphism data at and around the locus and from simulations that revealed that the Y139C mutation is an overdominant mutation under balancing selection that has entered our study area as a new mutation ($1/2N_e$) after the 1950s. Whether this means that the mutation arose de novo or arrived in from of an immigrant resistant rats cannot be established given the sampling and data. However, our simulations indicated that the mutation is unlikely very old (here as modeled as more than 700 generations) as this would have unlikely resulted in the pronounced selective sweep patterns we have observed and modeled. Nevertheless, solutions to this question require more empirical data. This study is the first to confirm the proposed classical example of balancing selection mode on the warfarin resistance gene by using molecular population genetic data and forward-time simulations covering the resistance causing Y139C amino-acid change as well as linked neutral SNP data. Our work is of significance in that it documents that 'classic selective sweeps' appear to less common in natural population than it was assumed (Pritchard, Pickrell, and Coop 2010; Hernandez et al. 2011). Similarly, even though the selection for warfarin resistance in rats has become a classical example of selection on a simple adaptive trait we showed that it is difficult, yet possible, to discern some of the details regarding the precise selective regime on that gene and mutation. Given some of the complexities encountered during this study, and others described in the following

chapters, we remain cautious in terms of the ability to interpret genome-scans for adaptive variants in an overly simplistic fashion.

Chapter 3

Network-guided GWAS reveals a polygenic architecture of warfarin resistance in the Norway rat

Abstract

Over decades, genome wide association studies (GWAS) have successfully identified markers associated with human disease or, in rare instances adaptive traits in natural populations of humans, plants or animals. A recent trend in the biomedical sciences is to improve traditional GWAS by incorporating gene-gene interaction network data. The power of this approach lies in the *a priori* knowledge that genes interact, and thus, the traditional approach to correct for multiple testing becomes obsolete or, less relevant. Here we use NetGWAS, as well as a Google PageRank algorithm that is modified as part of this thesis, to combine genotype-phenotype association measures with genetic network information to detect candidate genes underlying warfarin resistance in the Norway rat. Under this framework for analysis of genome-wide single nucleotide polymorphism (SNP) association genes that show significant levels of association under traditional analysis conditions are given high ranks if SNPs tagging genes that are neighbors in the gene-gene interaction network have high statistical support during the association study also. As gene scores are initially based on traditional measures for marker-trait association, new algorithms that consider gene-gene interaction network information adjust the original level of association based on known or suspected genetic interactions. Here we collected genome-wide SNP data for wild warfarin resistant Norway rats (*Rattus norvegicus*) from Germany with aim to investigate whether any other genomic regions than the vitamin K epoxide reductase subunit 1 gene (*Vkorc1*) in the rat genome display significant association with warfarin resistance, as was measured in our study by a

physiological blood clotting response test (BCR). Notably, warfarin resistance once was thought of as a simple Mendelian trait. However, here by assaying SNP polymorphisms across the rat genome we conducted a NetGWAS and identified 87 SNPs tagging genes that displayed significant association with the warfarin resistance trait. Thus, this study indicated that warfarin resistance has a polygenic architecture. Using our approach *Vkorc1* was significantly prioritized and attained the highest rank, as was expected and as we interpret as evidence for the power of our approach. However, we identified other genes whose annotations reveal compelling hints to previously unknown functions or pathways, many of which are in fact connected to the *Vkorc1* driven blood coagulation and bone developmental pathways, as well as biochemical reactions that take place in the endoplasmic reticulum where, in the case of hepatic cells, the vitamin K cycle proteins presumably are located. The population genomic analysis of a warfarin resistant population NW and a non-resistant population LH provided further support for some of these candidate genes. Notably, evidence for selection on candidate gene regions was found, indicating that the identified regions are not artifacts of population structure and drift alone. For example, we found elevated levels of linkage disequilibrium (LD) that are consistent with selective sweeps associated with these SNPs. Warfarin resistance in the Norway rat, once thought to be a classical example of a simple adaptive trait appears to be of a more complex architecture. Thus, other studies of natural populations likely will have to anticipate more complex genetic architectures also than is generally assumed, but as is shown in this chapter these can be reduced to some degree by adopting a NetGWAS approach.

3.1. Introduction

One important goal of genome-wide association studies (GWAS) is to dissect the genetic architecture underlying complex diseases in humans (Consortium 2007; Stranger, Stahl, and Raj 2011). To date (Dec, 2012), over 1,400 GWAS have identified around 8,000 SNPs related to human diseases and other traits since the first published GWA study in 2005 (Lucia A. et al.).

GWAS have been applied to search the ample human genome variation data for adaptive loci in humans (Hancock et al. 2010; Brown 2012). More recently and if the genomic tools were available, approaches reminicient of human GWAS have bee applied to plants and animals to identify genes underlying adaptive traits (Parker et al. 2009; Brachi, Morris, and Borevitz 2011).

As foreseen 15 years ago (Risch and Merikangas 1996), large scale association analyses have revolutionized the identification of complex diseases and traits. In most studies, however, a large number of variants are tested for their association with a phenotype individually, one-by-one. This results in major issues with significance testing in that a large number of associations between SNPs and trait could be false-positives. Mostly this issue has been addressed by developing methods to adjust P-values for multiple testing.

However, the value of utilizing results from molecular studies obtained in the laboratory has been recognized and increasingly implemented in GWAS in that genes are not assessed for their significance in isolation any longer but are evaluated with

regard to their connection to sets of genes that interact in gene-gene interaction networks (Cordell 2009).

Generally, there are two ways to consider such presumably 'known' gene-gene interaction information. First, one could incorporate known gene-gene interaction information to facilitate the identification of trait-associated sets of genes during GWAS (Morrison et al. 2005; Akula et al. 2011; Winter et al. 2012). Second, one could detect statistical interactions without applying any prior biological knowledge during GWAS and then, *a posteriori*, interpret statistically significant associations in light of known gene-gene interactions (Cordell 2002; Liu et al. 2011).

In this study we employed the first of these two strategies by combining a GWAS study with prior gene-gene interaction network information to prioritize the selection of candidate genes that are associated with warfarin resistance in the Norway rat. We explore if this NetGWAS approach holds promise in the study of natural populations as we note on the increasingly refined gene-gene interaction networks that are now available for many model species (Barabasi and Oltvai 2004; Barabasi, Gulbahce, and Loscalzo 2011).

Google's PageRank algorithm (Page et al. 1999), which helped to establish Google's success, has been successfully adapted to GO (Gene Ontology) network analysis (Morrison et al. 2005) and the analysis of gene-transcriptional networks (Winter et al. 2012). These successful adoption of the Google search algorithm in biological studies motivated us to perform similar such modification as suggested in the programming environments called GeneRank (Morrison et al. 2005) and NetRank

(Winter et al. 2012). These approaches couple SNP-trait association measures with gene-gene interactions that are mined from databases. This switch from an *a posteriori* evaluation of candidate genes to the implementation of algorithms that consider gene-gene interaction networks *a priori* is called NetGWAS.

Google's PageRank assigns a webpage a high rank if all the pages linked to it also have high ranks (Page et al. 1999). Similarly, genes with information on other genes that interact with it, and any information on associations with a phenotype, could be added to the ranking of candidate SNPs during GWAS. In such NetGWAS, the trait-association measure is assigned to each node in form of a gene score, and such gene scores are allowed to spread further depending on the network topology. Gene ranks are obtained by solving a convergence matrix of the ranking iteration. One desirable feature inherent in the algorithm is that highly relevant genes are prioritized and genes considered as noise in the particular biological context studied are given lower priorities. Thus, the approach might limit the ability to discover novel pathways as priority is given to known gene-gene interactions while rarely documented or unknown interactions are given a low priority. However, this approach might enable to at least focus on the systematic dissection of the genetics underlying a few seemingly complex traits. This would constitute main progress as the current state-of-the art analyses generally result in long lists of candidate genes that are virtually impossible to interpret with regard to the complex selection regime encountered by populations in nature.

In this study we applied NetGWAS to identify candidate genes involved in the resistance to the anticoagulant rodenticide warfarin in wild Norway rats (*Rattus norvegicus*). Warfarin causes lethal bleeding by impairing the recycling of vitamin K hydroquinone, which is essential for the gamma-carboxylation of blood coagulation factors and other vitamin K dependent proteins (c.f. Appendix 8, Figure 5.2) (Presnell and Stafford 2002; Stafford 2005).

Specifically, warfarin inhibits the activity of the vitamin K 2,3-epoxide reductase protein complex (VKOR) located in the endoplasmatic reticulum of hepatic cells, and one of the two subunits of this component is encoded by a vitamin K epoxide reductase subcomponent 1 (*Vkorc1*) gene (Presnell and Stafford 2002; Pelz et al. 2005; Stafford 2005).

Introduced in the 1950s, warfarin was a potent compound to control rat populations for 10-15 years, but in a surprisingly rapid manner resistance towards warfarin has become prevalent in Europe (Boyle 1960; Lund 1964; population and agriculture 1986) and US (Jackson and Kaukeinen 1972). Previous studies have identified *Vkorc1* as the primary gene that confers resistance, most notably by a Y139C substitution and possibly by other variations on the gene depending on the resistant populations and species under study (Rost et al. 2004; Pelz et al. 2005). The Y139C amino acid change at the amino acid sequence position 139 is a result of a non-synonymous mutation from A to G in exon 3. The mutation reduces the basal in vitro VKOR activity by ~50% (Pelz et al. 2005; Rost et al. 2009), but protect the VKOR from inhibition by warfarin.

We sampled Norway rats from an area in northwestern Germany (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003) where the Y139C mutation has spread and causes widespread resistance to warfarin in free-living rat populations. Data on warfarin resistance were available for the majorty of rats from previous studies that reported these (in form of the binary traits resistant/susceptible) as obtained from blood clotting response (BCR) tests (Kohn, Pelz, and Wayne 2003).

Warfarin is also the most popular oral antithrombotic drug used to reduce the incidence and risk of heart attack, stroke or thrombosis in humans (Gage et al. 2003; Geisen et al. 2005). Though being widely prescribed for about 60 years, warfairn remains a difficult drug to manage because of its narrow therapeutic window and dramatically varying dosages dependent on genetics, diet and environments (Kamali 2006). Few genes *VKORC1*, *CYP2C9*, *CYP4F2* have been established as dose predictors for individual patients. So far, the identified genetic factors predict ~40% of dose variance, and non-genetic factors (age, sex, weight, etc.) explain ~15% (Rost et al. 2004; Takeuchi et al. 2009). *GGCX, EPHX1* and other genes have been reported with controversial results from different studies in different populations (Wadelius et al. 2005; Pautas et al. 2009; Takeuchi et al. 2009). Nonetheless, incorporating genetic information greatly improved the warfarin dosage estimation and dramatically reduced health care costs (~ 1 billion annually, source: AEI-Brookings joint center for regulatory studies). It is now more commonplace to inform the warfarin dosing with the help of genetic information (Avery et al. 2011; UI-Health 2012).

Though recent genome scans in humans argued that there is no further need to add other genes than *VKORC1*, *CYP2C9* and *CYP4F2* to the list in near future, the 45% unexplained genetic variance in warfarin dosing does remains a puzzle (Cooper et al. 2008; Takeuchi et al. 2009). As warfarin becomes a poster child for the future of personalized medicine more genetic information might be crucial to more accurately assess the warfarin dosing requirements before prescription.

In rats *Vkorc1* is the only gene with solid evidence that documents its role in warfarin resistance. However, in this study we propose, and test, the hypothesis that persistent selection for warfarin resistance over many decades should have co-selected additional loci that either enhance the resistance trait, or, loci that act as modifiers in some other role (Figure 1.3). Moreover, preliminary investigations in our laboratory reported on rats that were warfarin-resistant according to BCR testing but without the known resistance mutation on *Vkorc1*. Finally, as indicated above, in humans more than the *Vkorc1* is involved in warfarin drug metabolism. Thus, the genetic architecture of warfarin resistance in the Norway rat likely is more complex than is currently assumed.

Warfarin resistance in rats is both advantageous and detrimental, depending on the presence of the poison in the environment. A substantial fitness cost of resistance incurs in form of vitamin K deficiency, sporadic hemorrhage, arterial calcification and reduced growth/reproductive rates as observed in previous studies (Markussen et al. 2003; Pelz et al. 2005; Jacob et al. 2012). Thus, the study of warfarin resistant rats might reveal interesting facets of an adaption in that both the

beneficial alleles as well as detrimental alleles might be detected, and their functional connections be revealed.

In this study, we aim to identify candidate genes involved in warfarin resistance using i) traditional GWAS, and ii) NetGWAS. First, we perform association tests at the level of single nucleotide polymorphisms (SNPs) collected on a microarray platform covering the rat genome. Then we map SNPs to genes and compute gene scores based on association measures. These association measures and gene scores will be evaluated in conjunction with gene-gene interaction network information as part of NetGWAS. The gene-gene interaction network is built based on the interaction information from STRING (functional protein association networks) database (Jensen et al. 2009). The candidate list will be evaluated in terms of the functions of genes and sets of interacting genes by using GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) annotations (Kanehisa and Goto 2000; Consortium 2008). Finally, we conduct a population genomic scan for extended haplotypes and linkage disequilibrium (LD) by comparing a resistant population NW with a nonresistant population LH.

3.2. Materials and Methods

3.2.1. Rat samples assayed on the SNP array

29 rats (17 males) of a laboratory strain (NW samples) were derived from wild rats sampled from an area with reported anticoagulant resistance in northwestern Germany (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). 21 of them

were resistant to warfarin according to BCR (blood clotting response) tests (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). We obtained 12 rats (9 males/3 females) from a non-resistant population LH (Location: LUDWIGSHAFEN) about 300km away from this resistance area (Appendix 1).

3.2.2. Genotype data

Genomic DNA of 41 rats was isolated from liver tissues using the classical phenol:chloroform extraction method. Genome scale SNP data of the rat samples were collected using the rat 10K array purchased from Affymetrix (http://www.affymetrix.com). DNA samples (each one with > 4ug in approximately 30 ul of 1X TE buffer) were sent to Baylor Genomic & RNA Profiling Core (http://www.bcm.edu/garp/) and Vanderbilt University for Quality control tests. The genome sciences resources microarray center at Vanderbilt University (http://array.mc.vanderbilt.edu/) conducted the genotyping experiment on our behalf. The genotypes were generated based on intensity files following data normalization and quality control using GTGS (Affymetrix GeneChip Targeted Analysis Software).

Genotypes were assigned based on the degree of the data from each assay belongs to a certain genotype cluster identified cross samples. Nine samples failed during genotyping as was judged by a call rate of less than 95%. Thus 20 NW samples and 12 LH samples were available for analyses. With 10,847 SNP sites from the Rat 10k array, after filtering 477 SNPs with failed genotype calls, we obtained 3,053 non-informative SNPs in wild rats, and 7,317 informative SNPs for rats from NW and LH population (including 54 SNPs with unknown Chromosome location).

We performed the following association tests and population genomic tests on all of the 7,317 informative SNPs and also on the dataset after filtering out SNPs with MAF (minor allele frequency) < 5% (6,506 sites). As results of both approaches are similar, we included all 7,317 SNPs such that we would not discard potentially interesting genes (Gorlov et al. 2008).

3.2.3. Genome-wide association analysis (GWAS)

The workflow of the genomic association analysis is depicted in Figure 3.1. First we performed the association analysis on the genotype data of informative SNPs using both genotype-phenotype chi-square test and Bayesian association analysis. The list of candidate SNPs was obtained following this traditional GWAS approach. Then, we mapped all SNPs on genes, and computed gene scores based on the association strength. Third, we combined the association information with the genetic networks to identify candidate genes based on the computed gene ranks. The network analysis and the traditional SNP association tests were complementary to each other. The analysis of function/pathway annotation was done based on Gene Ontology/KEGG.

3.2.3.1. Genotype-phenotype association test

For NW samples, association between warfarin resistance and genotype for each SNP was tested in PLINK (Purcell et al. 2007). A 2×2×K Cochran-Mantel-Haenszel (CMH) test was performed to control for the potential confounding factor sex (Kohn, Price, and Pelz 2008). Holm-Bonferroni step-down adjusted P-values, Sidak step-down adjusted P-values, Benjamini&Hochberg step-up FDR control, and

Benjamini&Yekutieli step-up FDR controls were computed. The HOLM corrected P-values were used to identify candidate SNPs under the null hypothesis H_0 of no causal SNPs are present in the data.

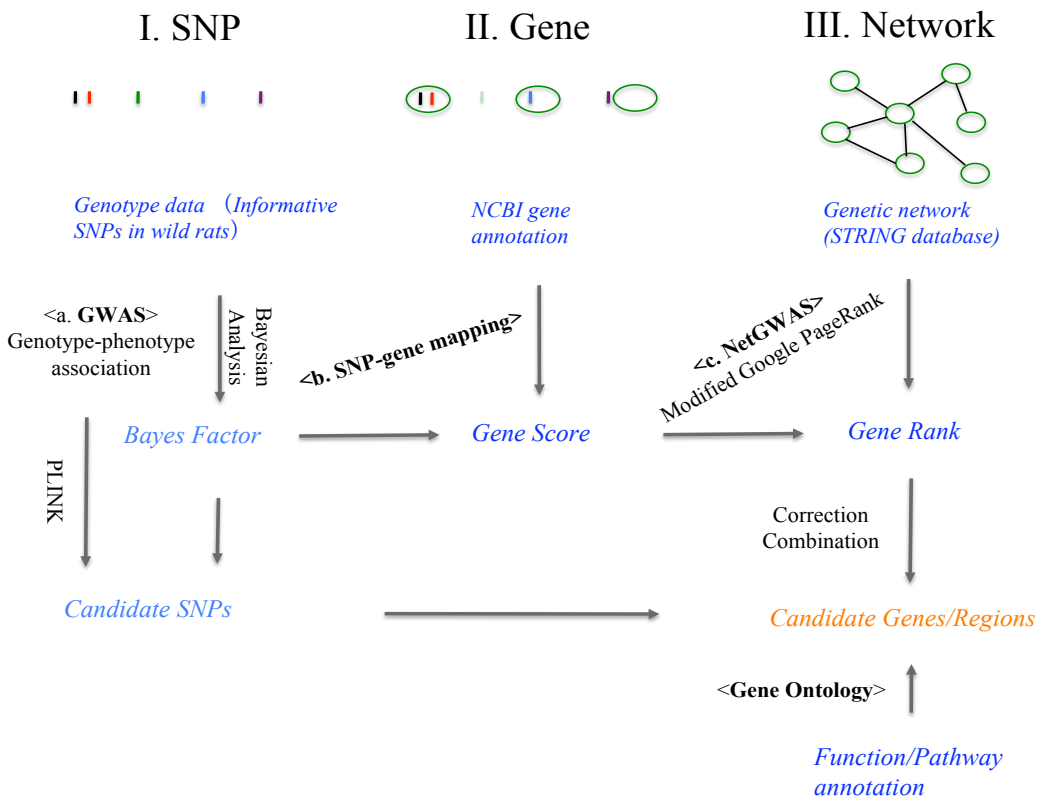


Figure 3.1 – The workflow of network-guided genomic association analysis.

3.2.3.2. Imputation-based Bayesian association mapping

We implemented the imputation-based association test for all rats and NW samples using software BimBam, using multi-marker LD model and a Bayesian regression approach (Scheet and Stephens 2006; Servin and Stephens 2007; Guan and Stephens 2008). The Bayesian regression method provided a Bayes Factor (BF) measuring the strength of genotype-phenotype association, and also generated a P-

value after permutation. First, with the integrated approach to perform imputation and computing Bayes Factors, we calculated BF1 for each SNP using prior D2 from (Servin and Stephens 2007) and averaging over σ_a (additive effect) = 0.1, 0.2, 0.4 and σ_d (dominance effect) = σ_a /4 as suggested (Servin and Stephens 2007; Stephens and Balding 2009). We also computed BF2 using $\sigma_d = \sigma_a$ to increase the weight on dominance model. The genome wide pattern of association was plotted with $-\log_{10}$ (P-values), $\log_{10}$ (BF1) and $\log_{10}$ (BF2) in R (Figure 3.6). We selected 82 SNPs with Holm adjusted P-values <= 0.05 or $\log_{10}$ (BF1) >= 1 or $\log_{10}$ (BF2) >=1 for future comparison.

3.2.3.3. SNP-gene mapping and gene score

We downloaded the *Rattus norvegicus* gene information of the NCBI build 4 (RGSC v3.4) from NCBI ftp (ftp://ftp.ncbi.nih.gov/, accessed March 2012). With the library file provided by Affymetrix Rat 10k array, we could assign SNPs to genes using the best SNP strategy with an choice of gene boundaries of ±50 kb to capture any potentially regulatory regions. SNPs could be assigned to multiple genes and these ambiguities were considered when evaluating significant SNPs at the later stages of the analyses. The boundaries of ±50 kb has been used in another gene-based association test software VEGAS (Liu et al. 2010). We also used the boundary choice of ±200 kb. We summarized the frequency distribution of the distance in megabases between SNPs and the nearest genes (Figure 3.2).

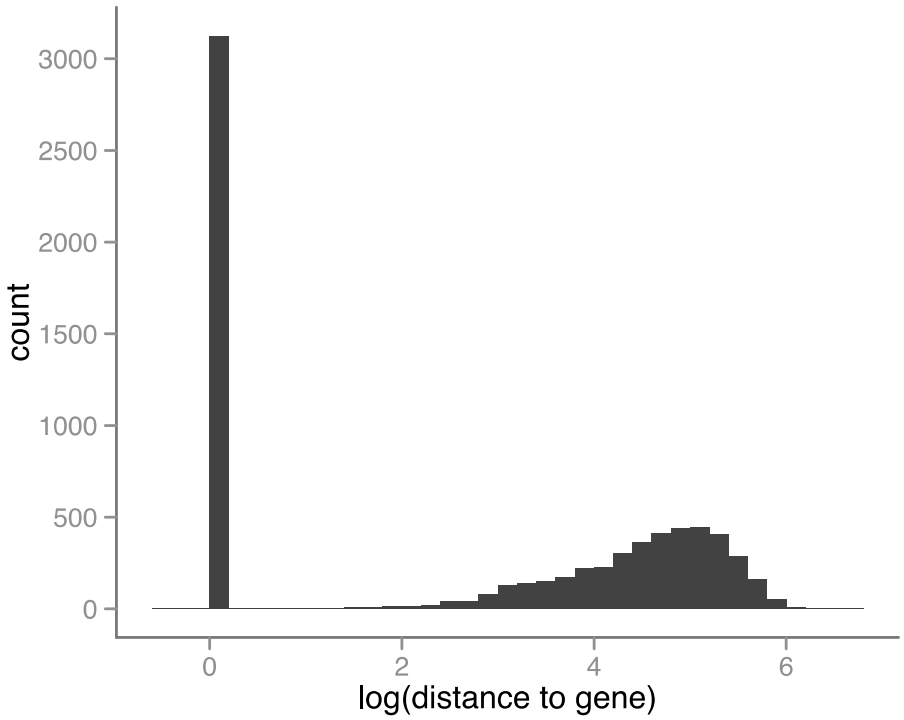


Figure 3.2 – The distribution of distance between SNPs and the nearest genes in log(bp).

According to the 'best SNP strategy', the association score of each gene was inherited from the SNPs with the highest Bayesian Factor that map within the genes, which measured the association strength. BF1 and BF2 were computed separately to provide a measure of significance under different genetic models. We did not compute traditional P-values of for association tests since these are lacking the power when measuring association strengths. To this end, we prepared the gene score files with either BF1 or BF2 measures each calculated for the gene boundaries of ±50 kb or ±200 kb to be used for the subsequent network-based ranking conducted. We removed the genes surrounding *Vkorc1* that were assigned with the resistance

mutation on *Vkorc1* since the flanking genes are known to be hitchhikers that map in the 30 Mb-spanning selective sweep region (c.f. Chapter 2).

3.2.4. Network-guided GWAS (NetGWAS) by a modified Google's PageRank algorithm

3.2.4.1. Genetic network

Incorporating the information of genetic interactions would facilitate the identification of candidate genes and functionally related additional genes. In gene-gene interaction networks each gene is considered as a node and the gene-gene interaction is considered as an edge. Here we considered an undirected such genetic network, in which the information was retrieved from the STRING database (v9.0) (http://string-db.org/, accessed July 2012). The downloaded *Rattus norvegicus* protein-protein interaction data contain information from 7 channels: neighborhood, gene fusion, coocurrence, coexpression, experiments, database and textmining. To incorporate the protein-protein interactions and potential functional relationships, we selected the links with scores ≥ 150 from experiments, database or textmining as suggested by the authors of the database. Then we converted the proteins into their corresponding gene names to obtain gene-gene relationships. We matched the genes from the gene score files with the gene-gene interaction information and built two networks. Network1 contains 4,846 genes from the above SNP-gene mapping result with ±50 kb boundaries, and 50,095 edges between genes. Network2 contains 9,382 genes from the above SNP-gene mapping result with ±200 kb boundaries, and 160,567 edges between genes. The Network2 used wider gene boundaries and thus

included more genes that match up with our SNP data, but Network2 might contain more false positive gene-interactions and noise due to cross-platform annotation issues. Thus we compared the results from both network analyses later. The distribution of degree of the gene-gene interaction network is plotted in Figure 3.3.

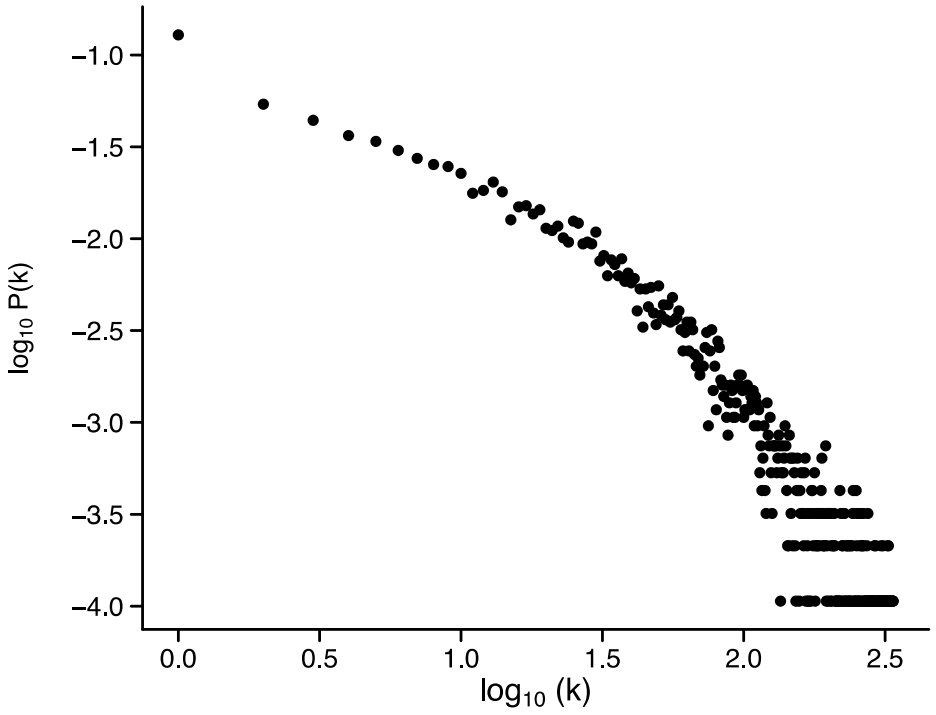


Figure 3.3 – The distribution of degree in the gene-gene interaction network

3.2.4.2. The modified Google's PageRank algorithm

Our NetGWAS combines the genotype-phenotype association information with the genetic network information to identify candidate genes by computing gene ranks. This network-ranking algorithm is adapted from the PageRank algorithm, which is key to ensure search quality of Google by giving users the desired information (Page et al. 1999). PageRank determines the significance of a page by the

importance of all the pages that link to that page. Similarly, in biological networks, the rank of a gene could be computed iteratively based on other genes that interact with it as well as based on their association with external traits. Here, we used a modified Google PageRank algorithm similar to one that has been used in gene expression analysis (Morrison et al. 2005; Winter et al. 2012) (Page et al. 1999):

$$R_i^m = (1-d)S_i + d\sum_{i=1}^{N}\frac{A_{ji}R_j^{m-1}}{\deg_i}, \qquad d \in (0,1)$$

Equation 3.1 – A modified Google's PageRank algorithm.

Where R_i^m is the rank score of gene *i* after *m* iterations; S_i is the score of gene *i*, here we used the association strength measured by Bayesian factor (BF1 and BF2 computed separately), but could be extended to any measure of interest; *A* denotes an adjacency matrix representing the gene network, in which $A_{ij} = A_{ji} = 1$ if gene *i* interact with gene *j*, otherwise $A_{ij} = A_{ji} = 0$; *N* is the number of nodes in the network; deg_j is the degree of gene *j*, which represents the number of interactions of gene *j*. *d* is a scalar parameter called damping factor, when $d = 0$, the impact of one node cannot be spread to others; whereas $d = 1$ means full influence from the network. The ranking vector *R* could be obtained by solving the linear system that equivalent to the convergence of iteration equation (1) (Page et al. 1999; Morrison et al. 2005; Winter et al. 2012):

$$(I - dA^T D^{-1})R = (1-d)S$$

Equation 3.2 – The convergence of Equation 3.1

Where I is the identity matrix; $D = diag(deg_i)$.

The choice of the damping factor d affects the ranking results. The Google damping factor of 0.85 might be higher than the suited setting for biological networks, since most genetic networks are much more sparse than the WorldWide Web (Clune, Mouret, and Lipson 2013). We tried two settings: first, $d = 0.3$ as evaluated in a previous study of transcriptional network (Winter et al. 2012) and second, we calculated the dynamic damping factor ddf_i for each node in our network according to Fu et al's suggestion (Fu, Lin, and Tsai 2006):

$$ddf = \frac{\deg_i}{\sum_{j|Aij=1} \deg_i}$$

Equation 3.3 – Dynamic damping factor.

Fu et al. suggested Equation 3.3 to dynamically use a damping factor in the ranking process (Fu, Lin, and Tsai 2006). However, applying the varying damping factor of different nodes is not expected to result in a solution of the eigenvector R for our undirected network. Thus, we calculated the average *ddf* across all nodes and used it as the damping factor d in Equation 3.2. We used MATLAB scripts to conduct the network rankings.

For each network (Network1 with ±50 kb boundaries and Network2 with ±200 kb boundaries) we obtained four files of gene rank score based on the calculated gene scores, BF1 and BF2, and by applying damping factors of 0.3 or the average *ddf*.

3.2.4.3. Evaluation of ranking results

To evaluate the gene ranking results, we compared the top 50 genes selected under four different scenarios:

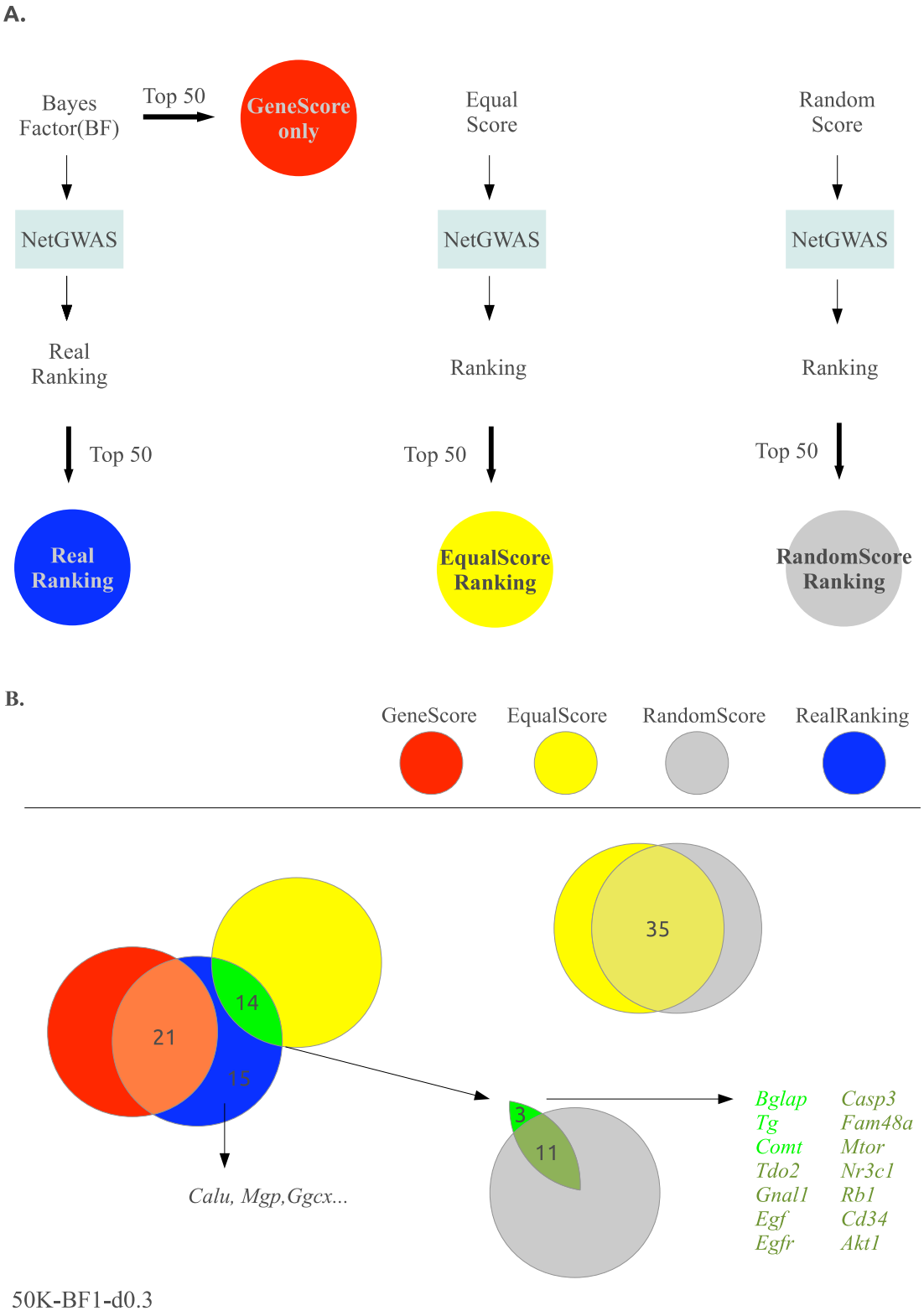
1) Gene score only. When d = 0, Equation 3.2 becomes R = S, i.e. gene ranks are only based on the assigned gene scores, which are the Bayes Factors.

2) Equal score. When d = 1, Equation 3.2 becomes $(I - A^T D^{-1})R = 0$, i.e. gene ranks are only based on network topology, which equates to an approach where the equal score is assigned to every gene.

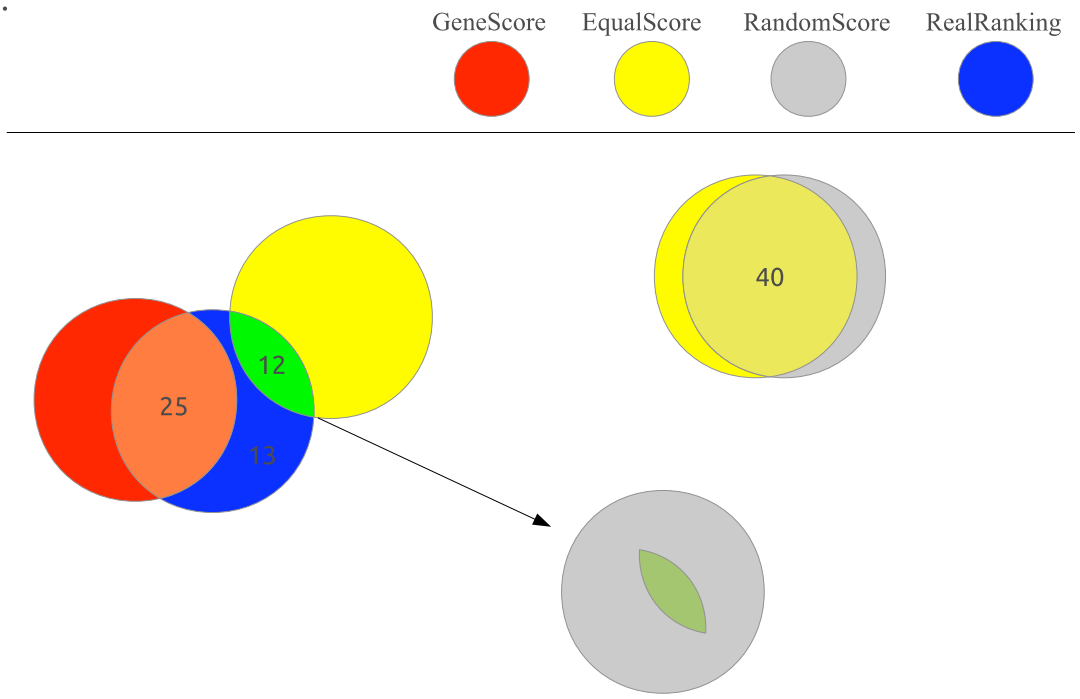
3) Randomized gene scores. To investigate whether some genes with randomly assigned scores could end up being selected as top 50 candidates, we randomly shuffled the association measures (BF2) among the genes in Network1 and Network2 10,000 times. We kept the network topology and then obtained 10,000 files of genes ranks based on such randomly assigned gene scores. The random probability for each gene score was calculated as the frequency of each gene (and gene score) selected as one of the top 50 candidate genes after 10,000 randomizations. Then the top 50 genes with highest random probabilities were selected for comparison. These random probabilities were also used for the following correction process.

4) Real gene rank score. The gene rank scores obtained as described above but modified to reflect the combined information from association analyses as well as network information in NetGWAS.

A comparison of results is shown in Figure 3.4B and C.

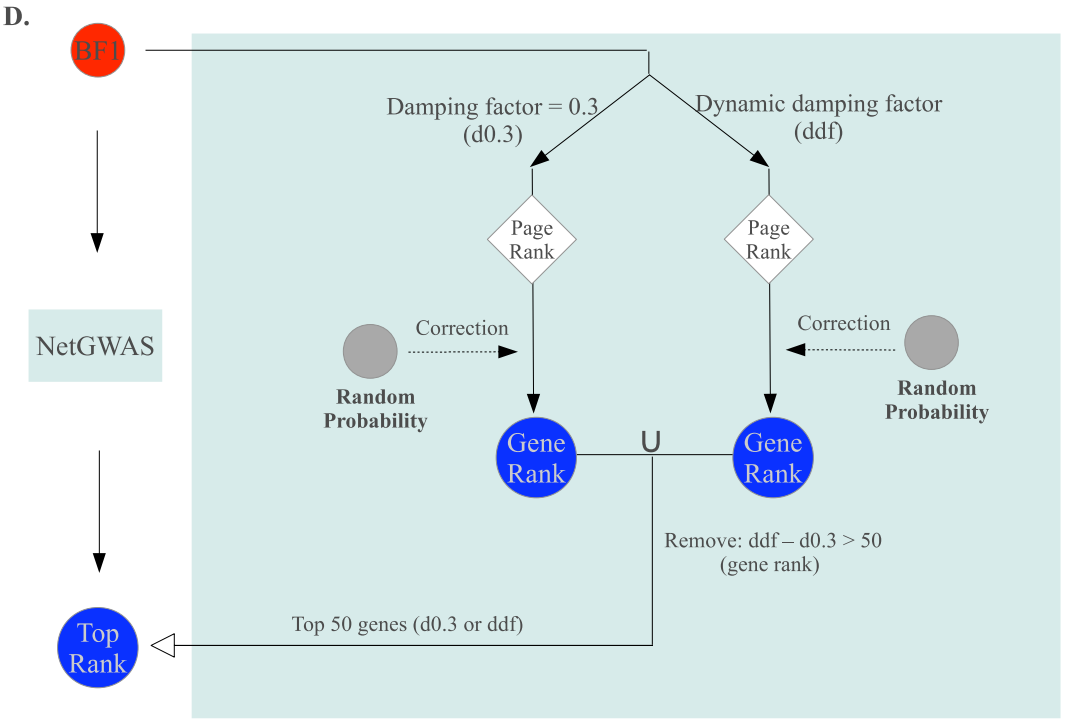

A.
Bayes Factor(BF)
Top 50
GeneScore only
Equal Score
Random Score
NetGWAS
NetGWAS
NetGWAS
Real Ranking
Ranking
Ranking
Top 50
Top 50
Top 50
Real Ranking
EqualScore Ranking
RandomScore Ranking
B.
GeneScore
EqualScore
RandomScore
RealRanking
21
14
15
35
3
11
Calu, Mgp,Ggcx...
Bglap
Tg
Comt
Tdo2
Gnal1
Egf
Egfr
Casp3
Fam48a
Mtor
Nr3c1
Rb1
Cd34
Akt1
50K-BF1-d0.3

C.



200K-BF1-d0.3

D.



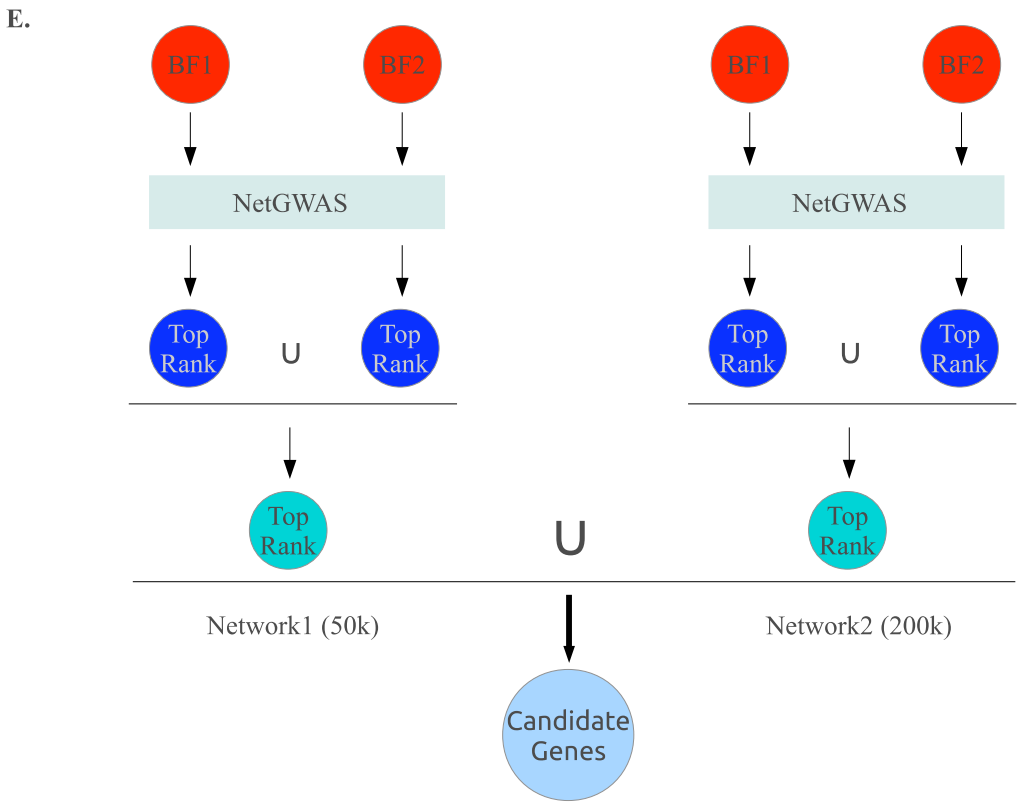


Figure 3.4 – The evaluation and correction of gene ranking results.

3.2.4.4. Correction methods applied to the gene ranking

As suggested by the above evaluation steps of genes, there are some genes to be expected in the data that have inflated high rank scores merely by virtue of network topology, rather than due to the biologically meaningful relevance to the trait under study. Thus, we preformed gene rank correction steps using the random probabilities calculated and by applying the two damping factors d=0.3 and *ddf*. First, we corrected the gene rank scores (vector R') by considering their random probabilities (RanProb):

$$R' = R(1 - \text{RanProb})$$

Equation 3.4 – Correct gene rank scores.

For example, *Vkorc1*, which has a low random probability, was ranked as a top candidate gene even after correction; whereas a gene with high random probability would be corrected to lower ranks.

Furthermore, we obtained gene rank scores based on the two damping factors: d=0.3 and *ddf* as suggested (Fu, Lin, and Tsai 2006; Winter et al. 2012). We also saw that the calculated average *ddf* is usually one order of magnitude smaller than d = 0.3. As mentioned earlier, the higher the damping factor the stronger is the effect of network topology on the final ranks of candidate genes. Without enough prior knowledge of the most suitable damping factor, at later stages of the candidate gene evaluation we closely examined those genes with big discrepancies in the rank scores obtained after application of the two damping factors. Especially some inflating effects of network topology could lead to a situation where genes with high rank score at d=0.3 can have low rank scores at *ddf*. As a complementary correction measure we used the random probabilities computed for each gene and its gene score after converting rank scores into ranks. We considered cases where $Rank_{ddf} - Rank_{d0.3} > 100$ as a large discrepancy worth investigating in more detail when evaluating candidate genes. Taking Network1 (±50 kb boundaries) and using BF1 as an example, there were 14 out of 50 top candidate genes that were also selected from the equal score case, i.e. only considering the network topology (Figure 3.4B). The correction process (using random probabilities and differences in damping factors) then removed 9 genes; if we applied stricter criteria by removing the genes whose $Rank_{ddf} - Rank_{d0.3}$

> 50, there were only 3 genes remaining in the intersection with the equal score case. Similarly, taking Network2 using BF1 as an example, 12 top genes, which also were be selected from the equal score case, were reduced to 1 gene after correction with $Rank_{ddf} - Rank_{d0.3} > 100$ and were all removed after correction with $Rank_{ddf} - Rank_{d0.3} > 50$.

File combination and candidate gene selection

As mentioned earlier, for each network we used two gene scores BF1 or BF2; and considering each gene score, we obtained the rank scores using two damping factors d=0.3 or the average *ddf*. In the case of BF1 calculations in our Network1 analysis we first corrected the gene rank scores by Equation 3.4 and then we removed the genes whose $Rank_{ddf} - Rank_{d0.3} > 50$. Subsequently we selected the top 50 genes from each file and pooled them together. Similarly for the BF2 calculations in our Network1 analysis we obtained the gene lists with corrected ranks after filtering out genes with large disagreements between the analyses done that applied the two different damping factors. We further pooled the gene lists from two gene scores (BF1 and BF2) and from two networks (Network1 and Network2) to yield the final list of 87 candidate genes (Appendix 4). The evaluation, correction and combination process is recorded in Figure 3.4D and E.

Besides combining the candidate genes from 8 (2×2×2) ranking files (supporting score: +1 if present for each file), we also checked whether the genes could be selected as top 50 based only on their Bayesian association scores (supporting score: +1 if present, for each gene score type BF1 or BF2), or whether

they were selected because of network topology (supporting score: -1 if present, for each network). Thus for each gene, we assessed whether they have multiple supporting evidences from different sources by summing up the supporting scores. We did this such that we have a large list of genes to work with when adding additional evidence (gene expression and population genomic data) and to be able to statistically evaluate groups of genes. We note that we will follow up on individual genes with additional genotyping and detailed analyses.

To investigate the interactions among the 87 candidate genes we drew a sub-network using the R package igraph (http://igraph.sourceforge.net/index.html) (Figure 3.7A). The bigger nodes correspond to genes with higher supporting scores. Some genes form a main cluster centered about *Vkorc1* (red nodes and edges). Some isolated candidate genes (blue nodes) were re-connected (blue edges) with the clusters if they are physically located near (< 5 Mb) the genes in network clusters. Orange nodes and edges represent small clusters with at least two interacting genes in network, which are separate from the main *Vkorc1* cluster in network.

The initial SNP association test suggested 82 candidate SNPs. Here we assigned these SNPs to above identified candidate genes based on their physical location using a < 2 Mb threshold (Appendix 4). Some candidate SNPs that cannot be assigned unambiguously to annotated genes are kept for future reference.

3.2.4.5. Function/Pathway evaluation

To evaluate the functions of candidate genes and the pathways they are involved in we performed GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis as implemented in DAVID (http://david.abcc.ncifcrf.gov, accessed Dec 2012). We submitted a background gene list with 9,382 genes from the larger dataset that emerged from the Network2 analysis and a candidate gene list (87 genes). The GO Fat category (particularly relevant to the vitamin K cycle) was disabled to avoid the overshadowing of specific functional terms such broadest of functional terms. KEGG and protein domain information (InterPro, PIR, SMART) were included in the analysis. The functional chart obtained listed 89 categories annotated from GO, KEGG or Protein Domain databases and the clustering report grouped similar annotations together, resulting in 37 clusters. After removing the functional clusters with enrichment P-values > 0.05, we obtained 19 clusters (Appendix 5).

We generated a GO network to facilitate the interpretation of functional relationships among genes discerned. In this GO network, nodes are genes, and two genes are connected by an edge if they have share similar functions (Figure 3.7B). Node size represents the supporting score for each gene. To examine the functional enrichment of each gene we also depicted the network in form of a heatmap where each cell represents the highest enrichment score of the category in each cluster for that gene (Figure 3.8). Genes are sorted based on their physical distribution along

chromosomes, and the left side bar uses the same color to represent chromosomal proximity (< 5 Mb) of genes in the network.

3.2.5. Population genomics

For the population genetic analysis of the SNP data we inferred haplotypes using fastPHASE while imputing missing genotypes (Scheet and Stephens 2006).

3.2.5.1. Haplotype structure

We computed two measures of haplotype identity: iHS and XP-EHH; both of them are based on Extended Haplotype Homozygosity (EHH). EHH estimated the probability of two randomly chosen haplotypes as being identical over a distance of X Mb to the locus under consideration (Sabeti et al. 2002). By definition, EHH depends on the choice of the relative distance to the locus. The integrated EHH statistic (iHS), is a standardized measure of EHH at a given locus of the ancestral allele relative to the derived allele (Voight et al. 2006). Thus large negative iHS scores suggest derived alleles under selection or the genetic hitchhiking of the loci. In contrast, large positive iHS scores indicate the presence of ancestral alleles. Treating each SNP as core SNP in turn, iHS summed over both directions of the EHH curve until 0.05 is reached for each site. Here we set the major allele frequencies observed in the non-resistant population LH as the frequency of the ancestral alleles (prior to warfarin selection) rather than determining ancestral and derived alleles based on outgroups (different species). The calculation and standardization of iHS are done with PERL scripts provided in WHAMM (Voight et al. 2006).

We also computed the XP-EHH statistic, which measures the cross population allele frequency differences between populations as the natural log of the ratio of iHS between two populations (Sabeti 2007), i.e. is a measure of population differentiation. The deviation from XP-EHH = 0 between resistant population NW and the non-resistant population LH can be interpreted in light of warfarin selection because population LH is fully warfarin susceptible whereas population NW is highly resistant and known to have been exposed to warfarin as part of experimental rodent control research. Limited by the SNP density of the Rat 10K array the resolution of these two measures of haplotype structure are accordingly limited. We assigned XP-EHH to the candidate regions (Appendix 4) and evaluated their values in light of the support they lend to our NetGWAS results. As a positive control we investigated the region on chromosome 1 where *Vkorc1* maps and both statistics provided clear evidence for selection at this region.

3.2.5.2. Linkage disequilibrium (LD)

We calculated the pairwise r-square values of SNPs and used haploview (Barrett et al. 2005) to assess LD blocks along each chromosome. Some of the candidate regions that featured large LD blocks are highlighted in Appendix 4. To study the rate of decay of LD with chromosomal distance we plotted the pairwise r-square values against distance in megabases separating any two SNPs for the whole genome data collected for the NW population (Figure 3.5).

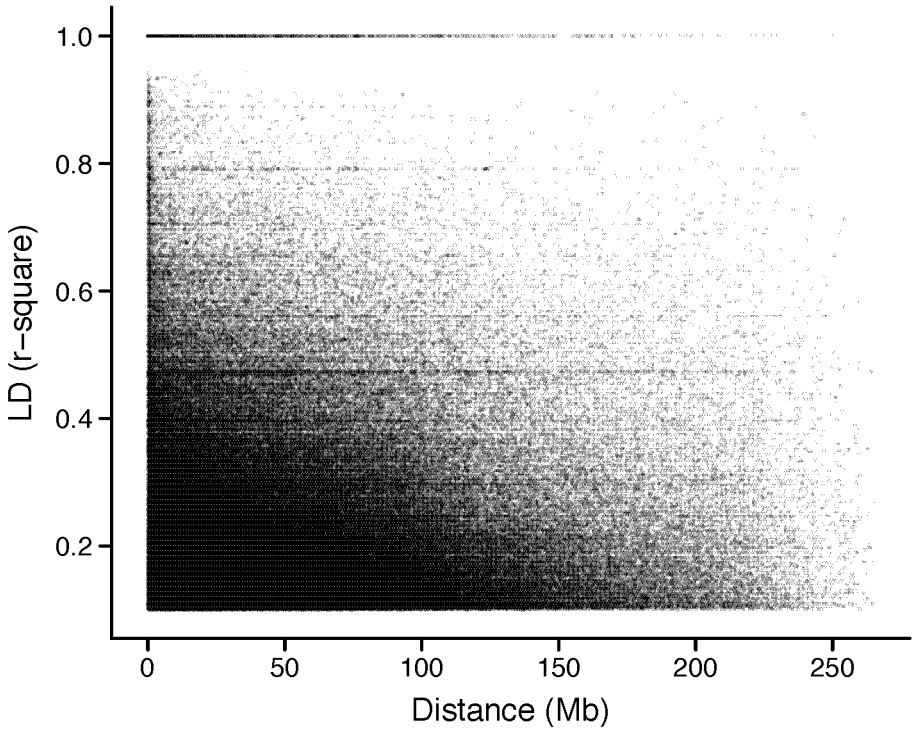


Figure 3.5 – Pairwise r-square against the distance between SNPs.

For the highlighted candidate regions with LD blocks in Appendix 4, we plotted the LD matrix of NW and LH samples using a modified R package snp.plotter (Luna and Nicodemus 2007). First, for the sweep region associated with the known resistant gene *Vkorc1* (172-202 Mb on chromosome 1), we surprisingly noticed that the LD blocks of NW non-resistant rats are stronger than resistant rats (Figure 3.9A and B). To understand this non-intuitive finding we performed forward-time population genetic simulations (see below for the parameter settings) for the *Vkorc1* sweep region using Python scripts in the environment of simuPOP (Peng and Kimmel 2005; Peng, Amos, and Kimmel 2007). We found that the LD blocks obtained for the sample of resistant rats in the population NW and for the corresponding sample of

non-resistant rats of the simulated population were plotted and compared (Figure 3.9C). For other interesting regions with LD block signals we plotted the pairwise r-square for rats collected from populations NW and LH, as well as for non-resistant resistant rats from population NW (Figure 3.9D-G). The comparisons between them potentially reveal the action of selection.

Forward-time simulation settings. We followed the parameter settings used during the simulation study described in Chapter 2. Briefly, population size was set as 1,000 diploid randomly mating individuals, and a recombination rate $r = 2 \times 10^{-3}$ per megabase per generation. Simulations were run to represent 100 generations corresponding to the time frame of 1965-1998 as estimated in Chapter 2. SNP sites: 100 SNPs located surrounding the Y139C mutation in *Vkorc1* (30 Mb) with their initial allele frequencies set according to the values in the control population LH. As suggested (c.f. Chapter 2), we assumed the adaptive variation on *Vkorc1* gene as a new mutation under balancing selection model. SNPs surrounding *Vkorc1* were hitchhiked with the adaptive variation. The selection coefficient *s* was 0.3 (estimated based on time-series data) with $t = 0.1$ as the fitness cost for the mutant homozygotes. The penetrance model for the adaptive SNP in *Vkorc1* was estimated as 0.069, 0.947 and 0.983 for wild homozygotes, heterozygotes and mutant homozygotes respectively. At the last generation, we obtained the genotype data for all the 100 SNPs and plotted their LD matrix for resistant and non-resistant samples respectively using the LD measure r-square.

3.3. Results

3.3.1. GWAS identifies candidate SNPs

As shown in Figure 3.1, genomic association analysis was first performed on 7317 informative SNPs. Then SNPs were mapped on genes, and gene scores are the association strengths of the assigned best SNPs. With a network perspective, we combined the association information with genetic networks to identify candidate genes by the computed gene ranks. Complementary to the traditional GWAS, the network-guided GWAS is a gene-based identification approach considering prior knowledge of gene-gene interactions and SNP association.

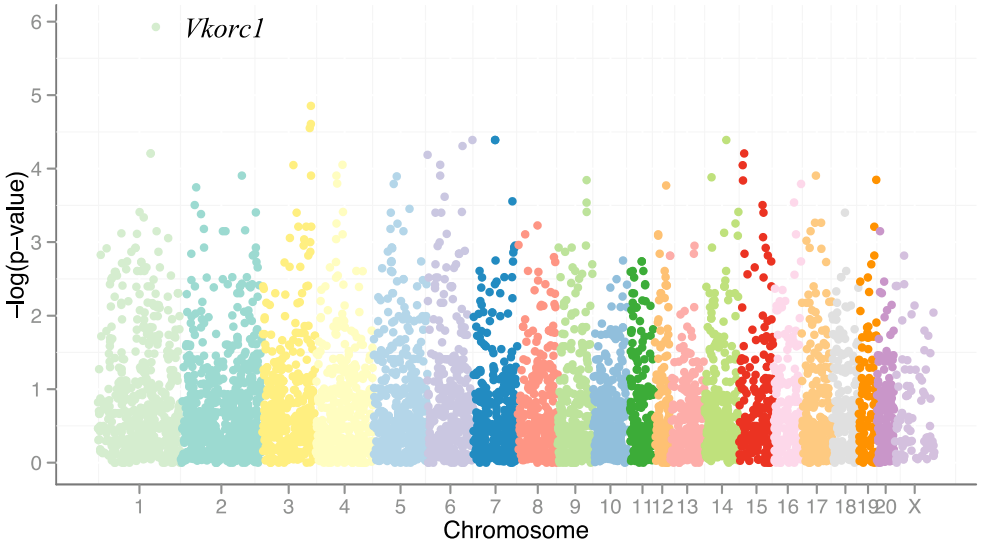
3.3.1.1. 82 top candidate SNPs detected based on association tests

29 rats from a resistant population NW and 12 rats from a non-resistant population LH were sampled from northwestern German and collected for rat 10k SNP array experiments (c.f. Methods and Materials). For the resistant population NW, when we performed the genotype-phenotype association test, an inflation factor $\lambda = 1.17$ showed some evidence for inflation of chi-squared tests (we don't pool NW and the non-resistant population LH together for association test because the high inflation factor of 2.9 implied the existence of population structure). So we do genomic correction and look at the genomic control adjusted significance value (GC) and the significance value corrected for multiple testing such as the step-down Bonferroni (HOLM) values. There are 31 SNPs with GC significance values < 0.05. After corrections for multiple testing only one SNP (Y139C) on the known resistant

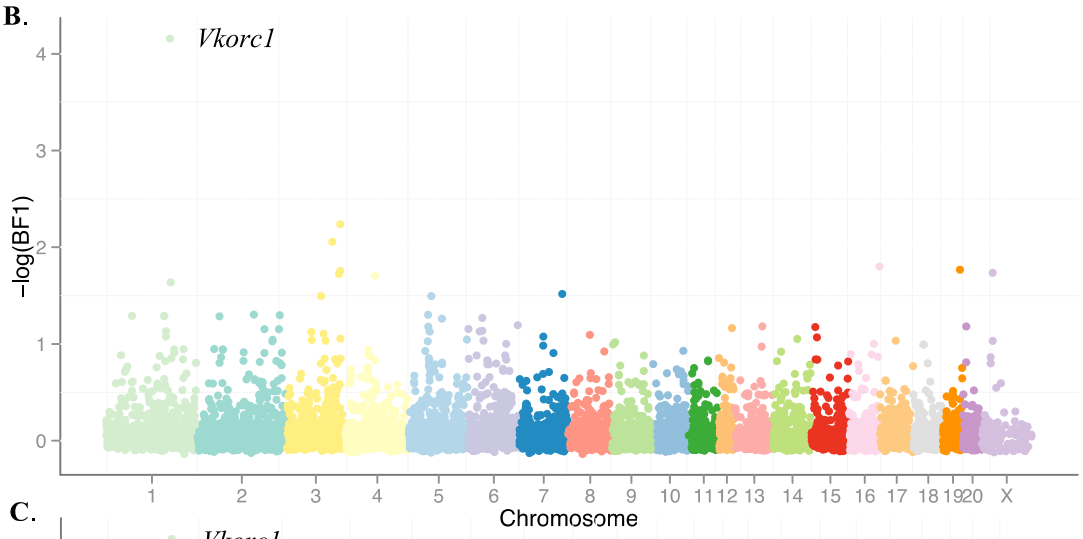
gene *Vkorc1* significantly associated with warfarin resistance (Figure 3.6A). The Bonferroni adjustment might be too conservative with the null hypothesis of no causal SNPs expected. Especially when we corrected for multiple testing based on genomic controlled values, even the known resistance SNP on *Vkorc1* gene showed no association.

Alternatively, Bayesian association analysis computed Bayesian Factor (BF) to measure association strength, which is similar to a likelihood ratio (Servin and Stephens 2007; Stephens and Balding 2009). The larger BF means stronger support for association (H_1) against no association (H_0). Thus $\log_{10}$ (BF) = 0 indicates not able to discriminate H_1 from H_0 (Stephens and Balding 2009). BF1 was computed assuming additive model and BF2 was assuming dominance model during the Bayesian analysis. Without prior knowledge of the genetic model for most regions, these two measures (BF1 and BF2) are considered together. Including the known resistance mutation (Y139C) on *Vkorc1*, there were only 3 SNPs with $\log_{10}$ (BF1) > 2 and 8 SNPs with $\log_{10}$ (BF2) > 2 (Figure 3.6B and C). We selected 82 SNPs (data not shown) with the $\log_{10}$ (BF1) ≥ 1 or $\log_{10}$ (BF2) ≥ 1 for future comparison with the results of a network-guided GWAS approach. Among the 82 SNPs, 15 SNPs are located in coding region. Two regions at chromosome 3 (157-158 Mb) and chromosome 5 (54-56 Mb) contain ≥ 3 SNPs.

A.


Vkorc1
−log(p−value)
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 X
Chromosome

B.


Vkorc1
−log(BF1)
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 X
Chromosome

C.

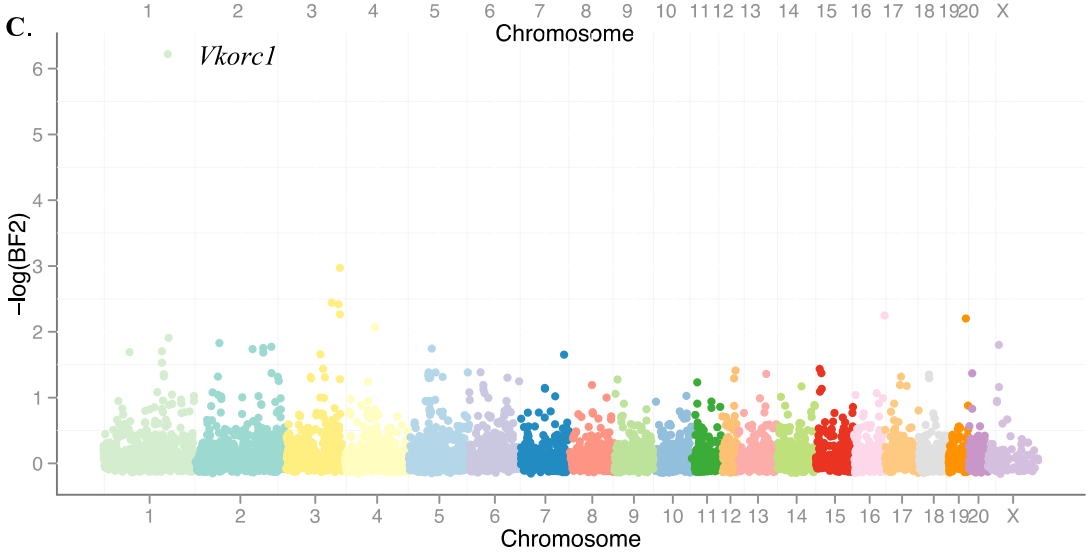

Vkorc1
−log(BF2)
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 X
Chromosome

Figure 3.6 – The genomic association signals (Manhattan plot). (A) The $\log_{10}$P-values of genotype-phenotype chi-square test (controlling for sex). (B) and (C) The $\log_{10}$BF1 and $\log_{10}$BF2 (Bayes Factors calculated in Bayesian association analysis assuming additive model and dominance model respectively).

3.3.1.2. SNP-Gene mapping and gene score

We move from SNP level to gene level by assigning SNPs to genes and computed the gene score based on the highest Bayesian Factor (BF) of SNP for each gene. As shown in Table 3.1 and Figure 3.2, about 1/3 (2,893 SNPs) of the 7,317 SNPs were located within genes, and 913 SNPs were in coding region. The ±50 kb boundaries as used in VEGAS (Liu et al. 2010) to capture surrounding regulatory regions were used during the mapping process. We also tried a wider gene boundaries of ±200 kb, which include more genes as well as potential noises. As one SNP could be assigned to multiple genes, we obtained 9,375 genes with scores using the ±50 kb boundaries and 19,775 genes with scores using the ±200 kb boundaries. The analysis results from these two gene score files would be combined later.

Table 3.1 – Statistics of SNP-gene distance (X).

All SNPs	Within gene (X=0 Kb)	0<X<1 Kb	1<X <10 Kb	10<X<100 Kb	0.1<X<1 Mb	≥1 Mb
7317	2893	253	826	1825	1455	11

3.3.2. NetGWAS identifies candidate genes by gene ranking based on a modified Google's PageRank algorithm

Genes are interacting with each other and the interactions are essential for performing and maintaining certain functions. With a network perspective, we could expand the traditional GWAS from isolated genes to interacting networks. We introduce NetGWAS, a modified Google PageRank algorithm, to combine genetic network information with gene-trait association measures. The PageRank algorithm, which basically ranks the importance of a webpage by the importance of other pages that link to it, is the key for the search quality of Google. PageRank have been successfully adapted to Gene Ontology network (Morrison et al. 2005) and transcriptional network (Winter et al. 2012), facilitating the identification of gene markers.

Here using NetGWAS, we integrate the information of phenotype association into genes as their features, which could spread in a gene-gene interaction network. Similar to the way of ranking webpages, genes interacting with more genes that are associated with resistance are given high ranks. Viewing in another direction, the association strength of one gene would spread to its neighbors and beyond depending on the network topology. NetGWAS naturally integrate association measures as gene features spreading in the network; thus this algorithm not only prioritizes genes associated with resistance but also reveals functional related genes missing in the initial association analysis due to low SNP density.

3.3.2.1. Gene-gene interaction (GGI) network

The gene-gene interaction (GGI) networks are based on the protein-protein interaction and text-mining information provided by the STRING database (http://string-db.org/). After filtering interactions with score ≤ 150 as suggested 344,481 interactions of 15,378 proteins remained for analysis. We matched proteins to genes, and then built two networks based on the genes from the gene score files with ±50 kb and ±200 kb boundaries. Thus Network1 analysis was based on 4,846 genes and 50,095 interactions and, since wider gene boundaries were applied when assigning SNPs to genes during analysis Network2 a larger number of genes (9,382) and interactions (160,567) were available for analysis. The network degrees were found to follow a scale-free distribution (Figure 3.3).

3.3.2.2. Gene ranking

Gene ranks are obtained by solving a linear system that on one side represents the gene-gene interaction network and on the other side represent the calculated gene scores (c.f. Methods and Materials). Besides network topology the damping factor *d* is a parameter defined to describe how far the association measures or other gene features would spread in the network. Google uses a damping factor of 0.85 for web searches. In biological networks, however, the damping factors should be set lower because genetic networks are sparse (Clune, Mouret, and Lipson 2013). We used d = 0.3 as suggested in a previous study of transcriptional networks (Winter et al. 2012). Also, we tried to calculate a dynamic damping factor ddf_i for each node in our network according to Fu et al's suggestion (Fu, Lin, and Tsai 2006) and used the

average *ddf* as the damping factor for the whole network. For our analysis Network1 (±50 kb boundaries) the average *ddf* is 0.03 and for the analysis Network 2 (±200 kb boundaries) average *ddf* is 0.02. Thus for Network1 and Network2, respectively, we obtained four files containing gene ranks based on BF1 and BF2 as gene scores computed using $d = 0.3$ or *ddf* as damping factors.

3.3.2.3. Evaluation and correction of gene ranking results

Based on the gene ranking results we investigated how different the results would be compared to those obtained based on the association strength measures calculated without considering any network information. Furthermore, there might be situations where some candidate genes are given high ranks because of their popularity (high degree) in the network instead of any association with warfarin resistance. If this were the case, then these genes would have high ranks under certain network topologies regardless of their gene scores. Thus, we evaluated the NetGWAS by comparing the ranking results under four different scenarios: (1) Only based on gene score; (2) Only based on network; (3) Randomized gene score; (4) Real gene ranking (c.f. Materials and Methods for details, Figure 3.4A). Taking BF1 in analysis Network1 as an example we found that the top 50 genes selected from the above four approaches display overlap (Figure 3.4B). For example, 21 genes selected based on both the gene association score (1) and gene ranking (4) emerged as candidate genes. However, 14 candidate genes from real gene ranking (4) analysis emerged also from the analysis based on equal scores (4) for all genes in the network, and thus should be treated with caution. Especially 11 of these 14 genes have high random probabilities

of being selected as top 50 in networks even when randomly shuffled gene scores (3) were assigned to them. This observation thus reminds us of the necessity of correcting the ranking results, foremost in light of the network topology.

As suggested by the above analyses there are some genes with inflated high rank scores due to network topology rather than trait relevance. Therefore we used random probabilities of being selected as top 50 genes with randomly shuffled gene scores and the difference between the two damping factors to correct the gene rank scores. As mentioned earlier, we obtained two files (with two damping factors: d=0.3 and *ddf*) for Network1 and BF1. Correction was applied to ranking results from each file, and then we selected the top 50 genes from the two files and pooled them together (Figure 3.4D). A similar correction process was applied to each gene score measure and each network before we pooled the lists from different analyses together to form a final list with 87 candidate genes (Figure 3.4E). In the process of combining candidate genes, we assessed whether genes had multiple supporting evidences from different sources by supporting scores (c.f. Methods and Materials, Appendix 4).

3.3.2.4. Candidate genes identified from GGI network

Comparison between the traditional GWAS results of candidate SNPs, 75 of the 87 candidate genes had top candidate SNPs nearby. Most (72 genes) of the SNP-gene distances are < 0.2 Mb and 16 genes contain the SNPs. 37 SNPs could not be matched to an annotated gene. Though most of these were with $\log_{10}$BF around 1-1.4, only one SNP (S695137) on chromosome 16 carried a relatively high association score ($\log_{10}$BF2 > 2).

Once we grouped the 87 candidate genes by their physical locations (within a 5 Mb distance) we observed 41 candidate regions. Considering the potential effect of a selective sweep on physically linked and proximal genes we highlighted those regions with multiple candidate genes located within them (Appendix 4). We observed that 48 candidate genes cluster in 10 regions with each cluster containing ≥ 3 genes. Conceivably, one or more genes located in each cluster are the causal genes under warfarin selection while the others appear as associated SNPs solely due to genetic hitchhiking effect.

As the primary resistance gene (Rost et al. 2004; Pelz et al. 2005), *Vkorc1* was ranked 1st in all the ranking files. Moreover, with high Bayesian association scores ($\log_{10}BF > 4$) and absence from the equal score ranking results, *Vkorc1* gained the highest supporting scores (12) calculated from the summary of results obtained from multiple sources. Did not observe a large region of candidate genes around *Vkorc1* here since we already manually removed these hitchhikers surrounding *Vkorc1* to decrease the noise in the data. Among the 87 candidate genes, there are 8 genes with the same supporting scores (12) as *Vkorc1* and a total of 45 genes with supporting scores > 5. Consideration of their Bayesian association scores resulted in are 3 genes with $\log_{10}BF1 > 2$ and 17 genes with $\log_{10}BF2 > 2$, in addition to *Vkorc1*.

The degrees of nodes, representing the number of network neighbors connecting any gene, are listed in Appendix 4. In the Network1 analyses the average degree of these 87 candidate genes is 29.8, which is somewhat higher than the mean

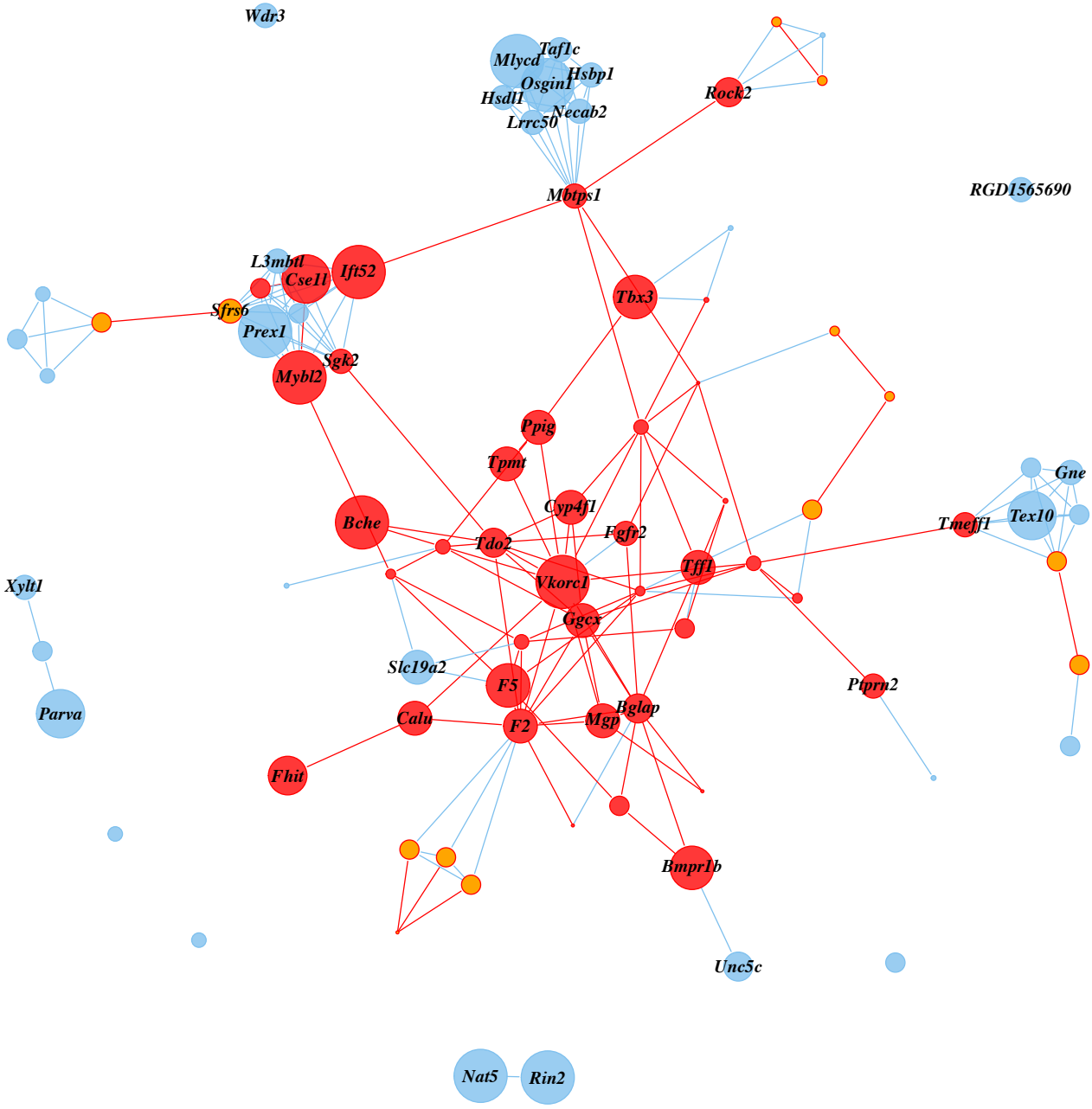
value of 21 computed for the entire network. In analysis Network2, the average degree of 58 candidate genes is 61, compared to 34 for the whole network.

Prior to the correction of the ranking results the average degree for the selected candidate genes was 83 and 155 for Network1 and Network 2 respectively. This indicates that indeed some candidate genes were selected due to their popularity (high degree) in the network, which may not necessarily reflect their relevance to warfarin resistance. However, after correction, most such inflated rankings appear to have been reduced appropriately. For example, previously selected *Akt1* has high node degree and high random probability; now is absent from the candidate gene list after correction.

We expect that recent and intense selection on adaptive variants would occur on genes with moderate degree, since hub genes with many neighbors tend to be conservative to changes and isolated genes with too few neighbors can hardly affect fitness. *Vkorc1* gene has moderate node degree of 9 neighbors in Network1 and 12 neighbors in Network2. Thus, this recently evolved mutation is predicted to have moderate number of effects on other genes but clearly has the potential to result in a multi-genic response to selection as well as the manifestation of a pleiotropic fitness cost by affecting important pathways (c.f. Chapter 2).

Incorporating network information for genomic association analyses tends to identify genes and their neighbors that are connected in the network if they also have relatively high association strengths.

A.


Wdr3
Taf1c
Mlycd
Hsbp1
Osgin1
Hsdl1
Necab2
Lrrc50
Rock2
Mbtps1
RGD1565690
L3mbtl
Cse1l
Ift52
Sfrs6
Prex1
Sgk2
Mybl2
Tbx3
Ppig
Tpmt
Gne
Cyp4f1
Bche
Fgfr2
Tmeff1
Tex10
Tdo2
Tff1
Xylt1
Vkorc1
Ggcx
Slc19a2
F5
Ptprn2
Parva
Calu
F2
Mgp
Bglap
Fhit
Bmpr1b
Unc5c
Nat5
Rin2

B.

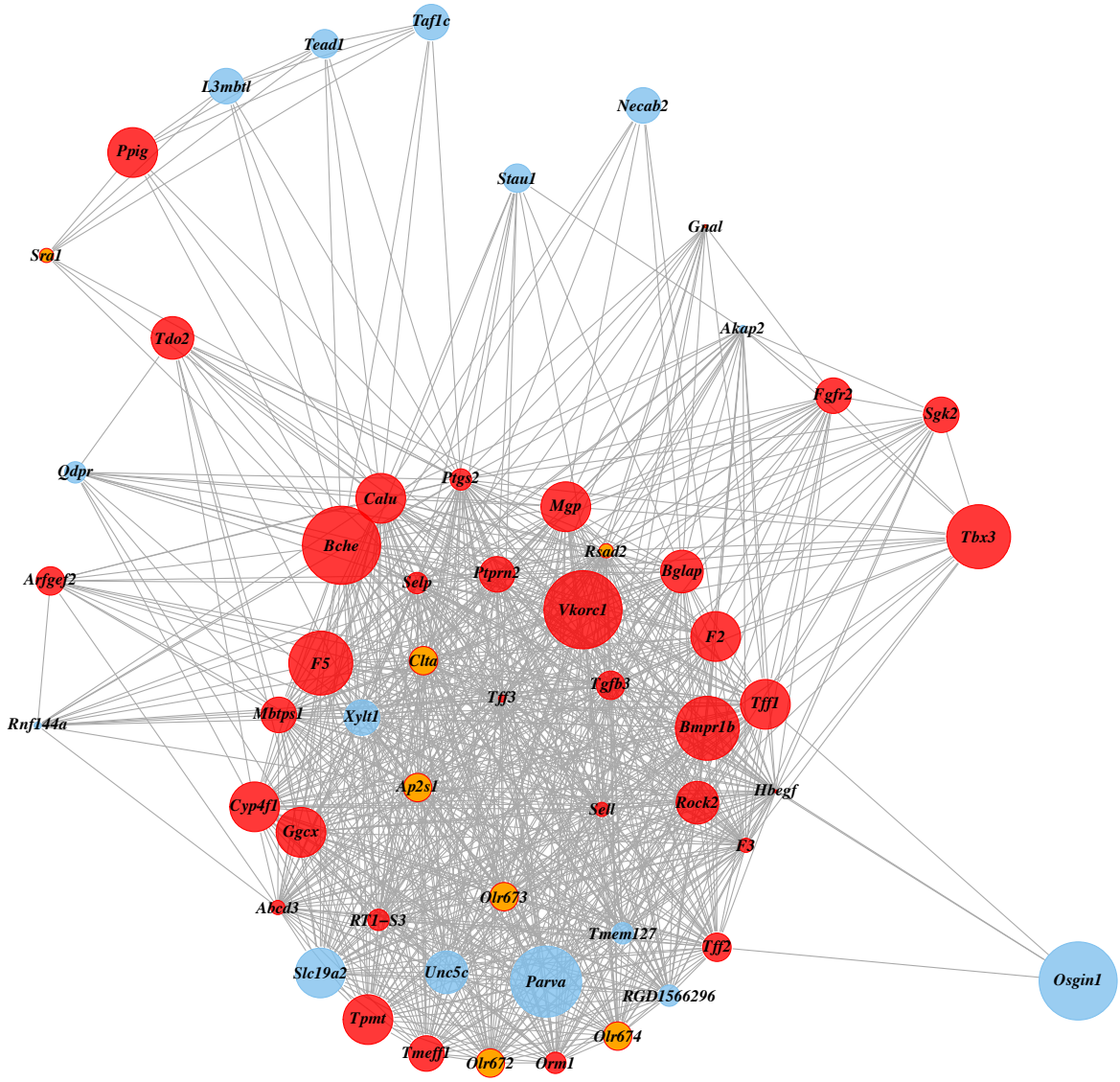


Figure 3.7 – The GGI network and GO for candidate genes. (A) GGI network. Nodes: genes; edges: gene interactions. Node size represents the support score of each gene. Genes (red nodes) belong to the cluster of interacted genes (red edges) centered with *Vkorc1*; isolated genes (blue nodes) and the genes belonging to other small clusters (orange nodes) are connected with the *Vkorc1* cluster by chromosomal proximity

(blue edges). (B) GO network. With the same color code for nodes, here the edges indicate that genes share similar functions based on function annotation analysis.

Here we examined whether these 87 candidate genes form clusters by building a sub GGI network showing the interactions among candidate genes. A main cluster centered about *Vkorc1* indicates that most candidate genes in fact could interact with the main resistance gene (Figure 3.7A, red nodes and edges). Bigger nodes are genes with higher supporting scores. Among the 40 genes in the main cluster, there are 12 genes that are directly connected with *Vkorc1* (*Ptgs2*, *Cyp4f1*, *Mgp*, *Calu*, *Ppig*, *Tdo2*, *Ggcx*, *Tpmt*, *F2*, *Bglap, Orm1, RT1-S3*). Separate from the main *Vkorc1* cluster, there are 34 isolated genes in terms of network interactions and 13 genes form 5 small but separate clusters.

However, if we also considered the physical location of candidate genes by connecting them in the network if the map within a 5 Mb chromosomal distance (blue edges), some isolated genes (blue nodes) and small clusters of genes (orange nodes) were then re-connected to the main *Vkorc1* cluster (Figure 3.7A). Only10 isolated genes remained separate from the main cluster, 5 of them located in 2 chromosomal regions. This representation allows us to explore the relationships among candidate genes in two dimensions: network interaction and chromosomal proximity. Further it enables us to distinguish those genes that may be of functional relevance in our study context from genes that merely are hitchhikers of nearby selective sweeps in candidate gene regions. For example, region 153 – 158 Mb at Chromosome 3 is composed of 9 candidate genes, and some of them reveal relatively strong association

signals. However, there is no prior knowledge about warfarin resistance and these genes. In Figure 3.7A we saw 5 genes (*Mybl2*, *Cse1l*, *Ift52*, *Sgk2, Arfgef2*) that are connected to the *Vkorc1* network cluster, and another 4 genes are just their physical neighbors. This observation connected this region on Chromosome 3 to the main *Vkorc1* cluster in network, and thus increases our knowledge about the potential candidate genes.

Overall, we obtained a 'warfarin resistance module' that mainly was composed of 40 candidate genes related to the vitamin K pathway, and some other candidate genes that are in chromosomal proximity to them (Figure 3.7A). This observation is consistent with the human 'disease module' which connects certain disease with a well-defined neighborhood of the interactome (Barabasi, Gulbahce, and Loscalzo 2011).

3.3.2.5. Candidate genes share similar functions

To evaluate the functions of the candidate genes, the GO (Gene Ontology) and KEGG (*Kyoto Encyclopedia of Genes and Genomes*) pathway analyses were performed in DAVID (http://david.abcc.ncifcrf.gov). For each of the functional (GO/pathway/protein domain) categories associated with candidate genes, the enrichment score was obtained by comparing candidate genes with the background gene list. Functional categories were grouped into 37 clusters by functional similarity. Keeping the clusters with at least one functional category with enrichment P-values ≤ 0.05, we obtained 19 functional clusters (Appendix 5).

To further explore their functional relationships, we built a GO network for candidate genes in which genes are connected if they share similar functions. As shown in Figure 3.7B, this GO network is densely connected (in average, each gene share function with 33 genes), which further implies the high functional relevance among candidate genes. More specifically, 40 candidate genes share at least one functional category with *Vkorc1* (GO/pathway/domain). To compare these results with the sub GGI network built from the list of candidate genes (Figure 3.7A), we used the same color codes for genes in this GO network. Most genes belong to the *Vkorc1* GGI network cluster (red) and constitute a dense cloud of high functional similarity (Figure 3.7B).

3.3.2.6. Summary of top candidate genes

In Table 3.2, we showed the top candidate genes with supporting score > 5 (all of the 87 candidate genes and more details on them are provided as Appendix 4). For each gene, we provide the top two functional terms. We also depicted the functional enrichment of genes across different functional clusters after assigning each gene with the highest enrichment score of the functional category in each cluster (Figure 3.8).

Vkorc1 is a known warfarin resistant gene and our approaches consistently ranked the gene 1st during both traditional GWAS and NetGWAS. According to Figure 3.8, *Vkorc1* is involved in 5 functional clusters (sorted from high to low enrichment scores): c6 - regulation of blood coagulation, c13 - positive regulation of

multicellular organismal process, c18-endoplasmic reticulum, c11 - response to organic substance and endogenous stimulus, c9 - disulfide bond.

Nine genes located in between 153-158 Mb on chromosome 3 exhibited relatively strong signals in terms of Bayesian association scores and supporting scores based on multiple ranking results (Appendix 4). *Prex1* in this region has the highest gene score ($\log_{10}$BF2 = 2.97) other than *Vkorc1*. But based on Figure 3.7A, *Prex1* was not connected to the *Vkorc1* cluster in the sub GGI network. Other two genes (*Ift52* and *Mybl2*) that map around 4 Mb away from *Prex1* also carry relatively high gene scores (2.42) and high supporting scores (12). Moreover, they are connected with the *Vkorc1* functional cluster in the network (Figure 3.7A). Thus this region is of great interest for further exploration, and we can start such explorations by examining the *Ift52* and *Mybl2* genes.

Another region located on chromosome 19 from 49.5-49.8 Mb caught our attention since 8 genes in this region have relatively high supporting scores. The two genes with the support score of 12 are involved in fatty acid metabolic process and growth factor activity. In this region, only one gene *Mbtps1* seems to connect with the *Vkorc1* cluster in the sub GGI network (Figure 3.7A). As shown in Table 3.2, *Mbtps1* shares the function of "endoplasmic reticulum" with other 5 genes including *Bche*, *Calu*, *Ggcx* and *Cyp4f1*.

Table 3.2 – Top candidate genes based on gene ranking (NetGWAS of SNP array I).

Gene	Support Score	Can SNP	Chr	GeneStart (Mb)	GS (logBF)	NetDegree	Function or GeneName
Parva	11	2	1	170	1.70	26	membrane
Xylt1	6	2	1	176	1.36	12	regulation of response to external stimulus;endoplasmic reticulum
Vkorc1	12	1	1	187	6.22	12	regulation of blood coagulation; positive regulation of multicellular process
Fgfr2	6	1	1	189	1.91	192	bone development; positive regulation of cell proliferation. (*FGFBP2*)
Bche	12	1	2	164	1.74	85	endoplasmic reticulum lumen;response to endogenous stimulus
Tdo2	7	0	2	174	0.12	323	oxidoreductase
Bglap	7	1	2	180	0.05	274	gamma-carboxyglutamic acid; bone mineralization
Wdr3	6	2	2	195	1.75	30	WD repeat domain 3
Unc5c	7	1	2	240	1.30	10	disulfide bond;extracellular
Bmpr1b	10	1	2	240	1.30	48	regulation of bone mineralization; reproductive developmental process
Ppig	8	0	3	52	0.19	55	intracellular organelle lumen
F2	8	1	3	76	0.21	105	gamma-carboxyglutamic acid; regulation of blood coagulation
Rin2	12	1	3	134	2.44	9	Ras and Rab interactor 2
Nat5	12	1	3	135	2.44	8	N-acetyltransferase activity
Sfrs6	6	1	3	**154**	2.42	48	serine/arginine-rich splicing factor 6
L3mbtl	6	1	3	**154**	2.42	1	intracellular organelle lumen
Sgk2	6	1	3	**154**	2.42	11	phosphorus metabolic process
Ift52	12	1	3	**154**	2.42	14	intraflagellar transport 52 homolog
Mybl2	12	1	3	**154**	2.42	45	myeloblastosis oncogene-like 2
Prex1	12	3	3	**158**	2.97	21	phospholipid binding
Cse1l	11	3	3	**158**	1.28	51	CSE1 chromosome segregation 1-like
Calu	8	0	4	56	0.56	51	endoplasmic reticulum lumen;cytoplasmic membrane-bounded vesicle
RGD1565690	6	1	4	88	2.07	5	GPRIN family member 3
Ggcx	8	0	4	106	0.23	24	disulfide bond;endoplasmic reticulum
Mgp	8	0	4	174	-0.03	118	gamma-carboxyglutamic acid; bone mineralization
Gne	6	1	5	**61**	1.30	37	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase
Tex10	11	1	5	**65**	1.75	4	testis expressed 10

Tmeff1	6	1	5	**65**	1.75	47	domain:EGF-like; disulfide bond
Rock2	7	2	6	41	1.27	84	Rho-associated coiled-coil containing protein kinase 2
Ptprn2	6	1	6	144	1.25	45	membrane-bounded vesicle; cytoplasmic membrane-bounded vesicle
Cyp4f1	8	0	7	14	0.30	39	endoplasmic reticulum; membrane
Tbx3	10	1	12	38	1.41	25	skeletal system development; reproductive developmental process
F5	10	1	13	**80**	1.36	52	cytoplasmic vesicle part;wound healing
Slc19a2	8	1	13	**80**	1.36	10	integral to membrane
Fhit	9	2	15	17	1.37	161	fragile histidine triad
Tpmt	8	0	17	24	0.14	20	integral to membrane
Hsbp1	6	1	19	**50**	2.20	7	heat shock factor binding protein 1
Necab2	6	1	19	**50**	2.20	1	N-terminal EF-hand calcium binding protein 2
Mlycd	12	1	19	**50**	2.20	34	fatty acid metabolic process
Osgin1	12	1	19	**50**	2.20	5	growth factor activity
Mbtps1	6	1	19	**50**	2.20	140	endoplasmic reticulum;Golgi apparatus part
Hsdl1	6	1	19	**50**	2.20	1	oxidation reduction
Lrrc50	6	1	19	**50**	2.20	6	cell morphogenesis
Taf1c	6	1	19	**50**	2.20	57	intracellular organelle lumen
Tff1	8	1	20	10	1.37	54	secreted;growth factor activity

SupportScore: genes' support score based on multiple ranking results. Bold gene names are genes with highest support scores.
CanSNP: the number of matched candidate SNPs from traditional GWAS.
Chr and GeneStart (Mb): genes' position on chromosome.
GS(logBF): gene score based on Bayes Factor ($\max(\log_{10}BF1, \log_{10}BF2)$ here).
NetDegree: the number of other genes connected to each gene in network (Network2 here).Function or GeneName: the description of GO terms (KEGG or protein domain category if no GO) within the cluster with the top two enrichment score for each gene;

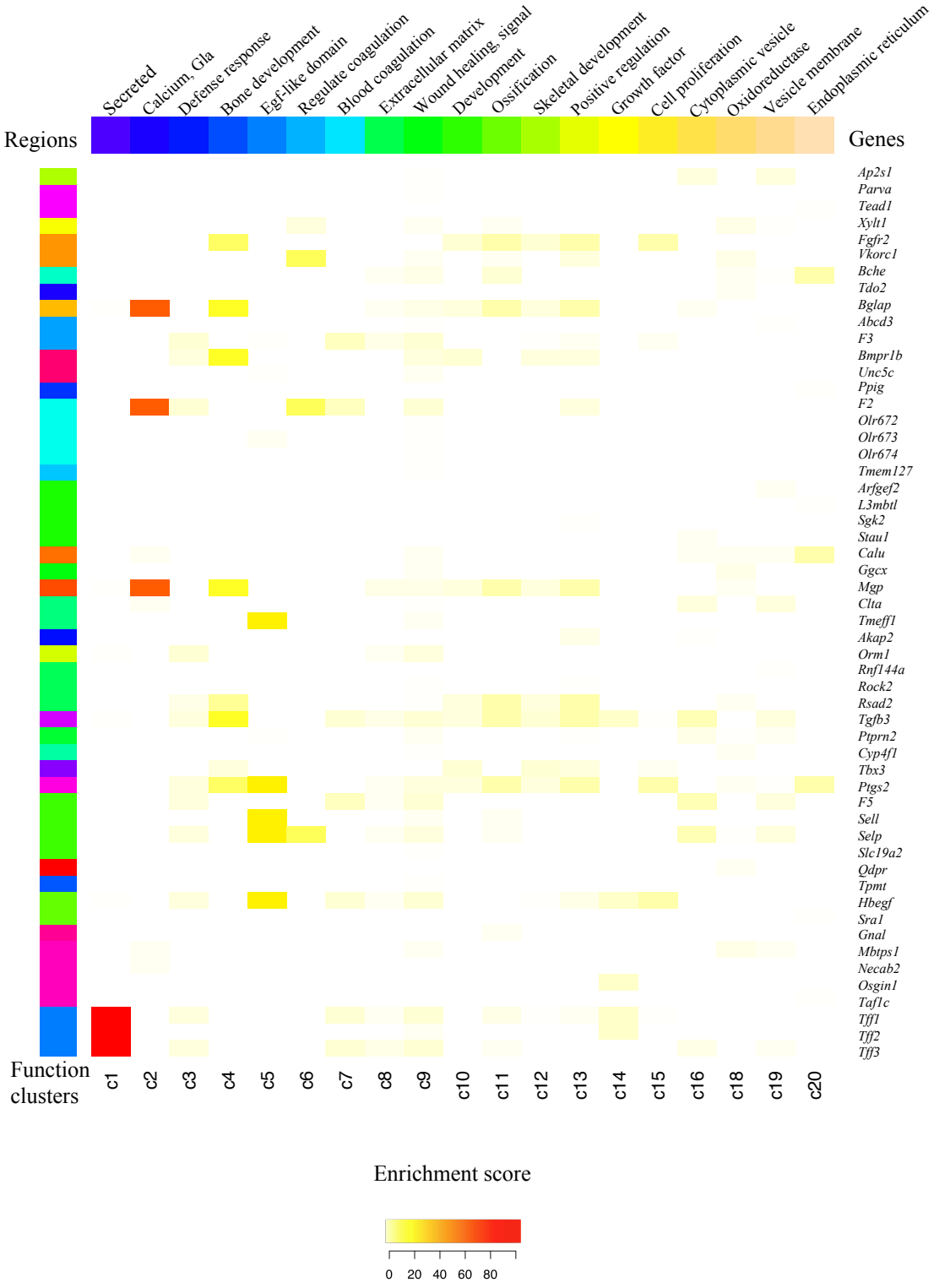


Figure 3.8 – The enrichment anaysis of functional categories. Left sidebar use color to separate candidate genes from different regions.

Bche carries the highest support score of 12 as *Vkorc1*. *Calu* and *Ggcx* are the genes involved in vitamin K cycle. Warfarin inhibits VKOR (encoded by *Vkorc1*) activity, and thereby impairs the recycling of vitamin K hydroquinone, which is an essential cofactor for gamma-carboxylation of downstream vitamin K dependent proteins (Presnell and Stafford 2002; Pelz et al. 2005; Stafford 2005).

Ggcx, which encodes the gamma-glutamyl-carboxylase complex (GGCX), is the other important cofactor in the gamma-carboxylation system (Markussen et al. 2007b). *Calu* has been shown to regulate the vitamin K-dependent gamma-carboxylation system (Markussen et al. 2007b) and competes with warfarin for the binding-site in the VKOR complex (Markussen et al. 2007a). *Cyp4f1* gene belongs to the cytochrome P450 gene family; most *CYP450* gene members encode enzymes functions in metabolism of small molecules and xenobiotic compounds (Thomas 2007; Pautas et al. 2009). In addition to *VKORC1*, the other two gene markers (*CYP2C9* and *CYP4F2*) used in human for dosage prediction are also members of this family (Rost et al. 2004; Takeuchi et al. 2009). *CYP2C9* metabolizes the S-form warfarin, and *CYP4F2* is believed to catalyse hydroxylation of vitamin K1 (McDonald et al. 2009). In the rat, *Cyp4f1* gene is the orthologous gene to *CYP4F2* in human.

As shown in Table 3.2 and Appendix 4, some vitamin K dependent proteins were selected out as candidate genes (also c.f. Appendix 8, Figure 5.2), such as *F2*, *F5*, *Bglap* and *Mgp*. Both *F2* and *F5* are blood coagulation factors activated via gamma-carboxylation system, which thus are directly affected by VKOR activity and warfarin (Stafford 2005). Some other vitamin K dependent proteins *Bglap* (bone gamma-

carboxyglutamic acid-containing protein) and *Mgp* (matrix gla protein, calcification inhibitor) play roles in vessel calcification and bone mineralization (Suttie 1993; Danziger 2008). A previous study did observed arterial calcification in resistant rats with reduced activity of VKOR (Kohn, Price, and Pelz 2008). Thus, it is interesting to encounter these vitamin K dependent genes and other such genes (*Bmpr1b* and *Fgfr2*) involved in bone development as candidate genes related to warfarin resistance in this NetGWAS. *Fgfr2* (fibroblast growth factor receptor 2) is 2 Mb away from *Vkorc1* on Chromosome1, so it might be hard to distinguish true association from false association due to the selective sweep at *Vkorc1*. Interestingly, a recent genomic scan in humans for warfarin dose related genetic variants revealed *FGFBP2* (fibroblast growth factor binding protein 2) gene has a relatively high genotype-dose association in addition to *VKORC1* and *CYP2C9* (Cooper et al. 2008). This finding reminds us of the potential importance of *Fgfr2* gene even though it might be under the sweep effect on the resistance gene *Vkorc1*. We noticed that *Bmpr1b* is also related to reproductive development, which might be a coincidence or an interesting lead as resistant rats have reduced reproduction and growth rates. Previous studies also suggested that the effect of MGP on calcification is determined by the relative amounts of MGP and the bone morphogenetic protein-2 (BMP-2) (Zebboudj, Shin, and Bostrom 2003). Here we observed the gene *Bmpr1b* in our candidate list and the gene is a bone morphogenetic protein receptor.

In addition, there is one more region at 60 - 65 Mb on chromosome 5 covering multiple (6) genes (Appendix 4). In this region, the *Tex10* gene has a high support score, and it is a testis-expressed gene, which suggests us to further explore its role in sex

difference related to anticoagulant resistance, as male rats seem to be more sensitive to rodenticides (Kohn and Pelz 1999).

We noticed that some candidate genes have low gene scores (with logBF < 1). They were prioritized, however, based on their interactions with genes associated with warfarin resistance. For example, the *Orm1* gene has a low (logBF1 = 0.02) association strength measured by Bayes Factor but it is interesting to retain it here in the candidate list because a previous study on warfarin pharmacokinetics revealed that warfarin is bound to the proteins encoded by *ORM1* and *ORM2* during its transportation from stomach to liver in human (c.f. Appendix 8, Figure 5.2) (Wadelius et al. 2007). These genes would probably be missed in the traditional GWA study without network information, especially when SNP density is low.

3.3.3. Population genomic analysis supports 6-8 out of candidate regions

Selection would leave differentiated genetic signatures between populations under different environments. Here we compared the population genetic signals of the resistant population NW with the non-resistant population LH. Limited by the low density of SNP data, we used these signals as supporting evidence for above candidate gene regions.

3.3.3.1. Measures of haplotype structure detect recent selection

Extended haplotype homozygosity (EHH) has first been developed to detect recent selection by assessing the decay of LD along the chromosome (Sabeti et al. 2002). Then other two measures of haplotype structure were developed based on it. By comparing the integrated EHH between ancestral and derived alleles, large positive or

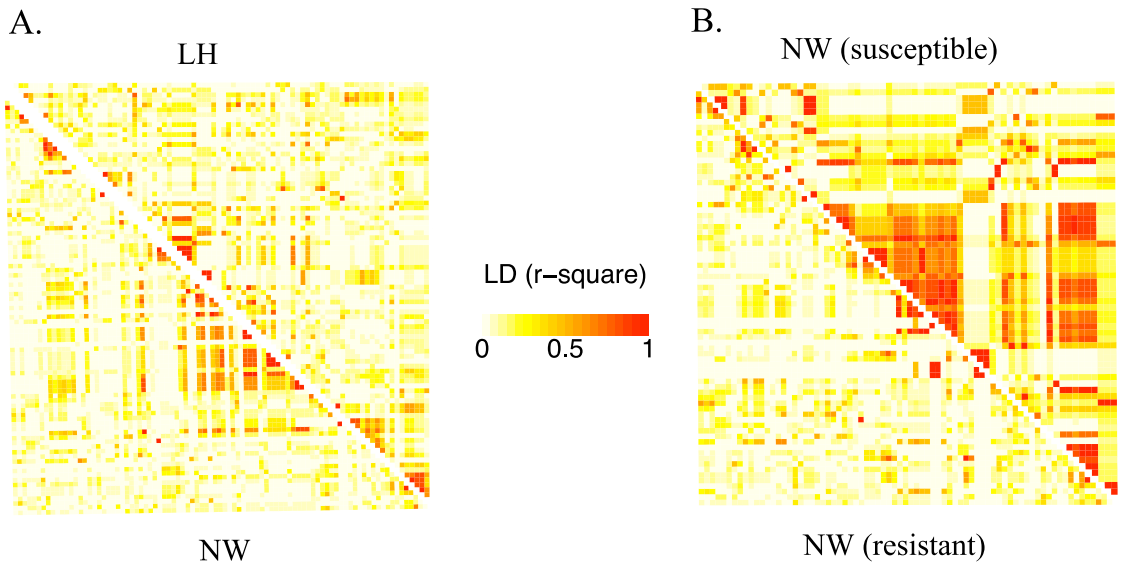
negative iHS score indicate unusually long haplotype carrying ancestral or derived alleles respectively (Voight et al. 2006). We calculated iHS scores and found 27 SNP with |iHS| > 2 in the NW population; 17 of them could be matched to 8 candidate regions (Appendix 4). However, iHS could not be calculated around the known resistant gene *Vkorc1* because of low SNP density. We also computed the cross population EHH (XP-EHH), which is the natural log of the ratio between the iHS score from LH and NW population (Sabeti 2007). Large negative XP-EHH thus suggests selection in the NW population. In total, we found 12 SNPs with XP-EHH < -2; but only 2 of them could be matched to candidate regions (Appendix 4). One peak of both the iHS and XP-EHH measures is the chromosome 12: 34 – 38 Mb region, including the *Tbx3* gene, which is involved in reproductive developmental process and mediated by bone morphogenetic protein (BMP) signals (Chen et al. 2009).

3.3.3.2. Blocks of linkage disequilibrium reveal signals of selection

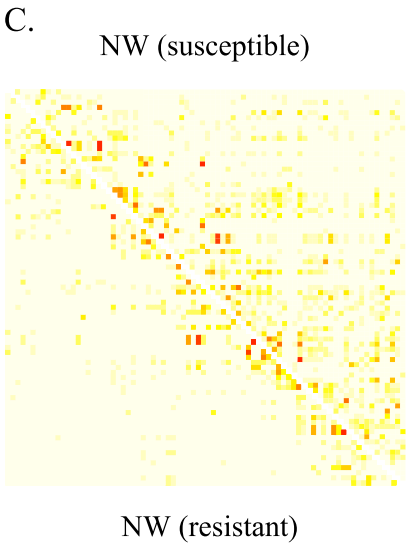
To check the pattern of LD decay over distance, we plotted the r-square values along chromosomes in both the NW and LH populations (Figure 3.5).

Selection usually creates LD blocks different from the background, thus regions carrying adaptive genetic variants should be detectable by searching for LD blocks. In the software Haploview (Barrett et al. 2005) we compared LD patterns using the pairwise r-square matrix of each chromosome for both the NW and LH populations. LD blocks at 6 regions were found and marked besides the candidate gene regions (Appendix 4). There are 3 LD blocks (Chr1: 249 Mb, Chr6: 72 Mb and Chr12: 21Mb) that were recognized in Haploview but that cannot be matched with any of candidate gene regions, which might

be interesting or simply due to demographic effects. Below we focus on LD block regions where candidate genes map, and these are the *Vkorc1* region, *Ggcx* gene, *Tex10* gene region and the region on chromosome 3: 153-158 Mb that have support from LD analysis.



Vkorc1 sweep region (Chromosome 1: 172 – 202 Mb)



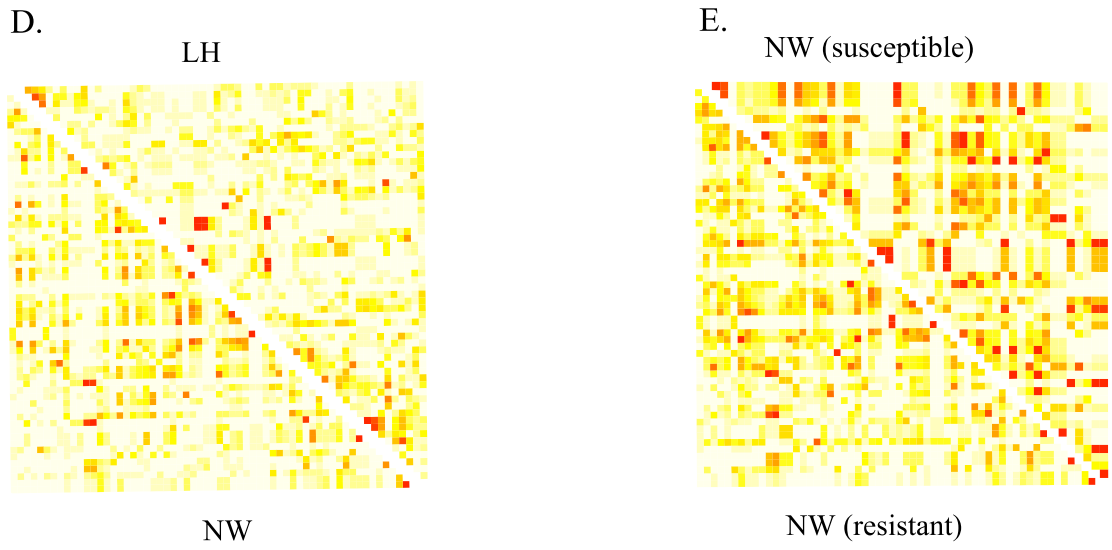


Figure 3.9 – The linkage disequilibrium blocks in resistant population NW and non-resistant population LH (A and D); LD blocks in resistant and susceptible samples within NW population (B and E). Simulated LD pattern in resistant and non-resistant rats (C).

First, for the known resistance gene *Vkorc1*, LD blocks surrounding it (172-202 Mb on Chr1) were observed in the resistant population NW but not in the control population LH (Figure 3.9A). But surprisingly we also found that the NW non-resistant rats have stronger signals of LD blocks than resistant rats (Figure 3.9B).

To assess whether this observation is merely a special pattern in NW population or a generally useful signal, we performed forward simulations with the selection coefficient $s = 0.3$ as well as other parameters as estimated in a study of *Vkorc1* sweep patterns (c.f. Chapter 2). After 100 generations we plotted the pairwise LD based on simulated data within resistant and non-resistant rats respectively. We did observe that LD signals of non-resistant rats are stronger than in resistant rats in the simulated

population after selection (Figure 3.9C). The differences of LD signals between resistant and susceptible rats were not as strong as we observed in Figure 3.9B, since 1,000 rats in population were allowed with random mating in the simulation, which might be a larger population size than the true effective population sizes of isolated demes of free-living rats.

For other candidate regions, such as the region at Chr3 from 141 to169 Mb, we compared the LD blocks between the NW and LH populations, and also between the NW resistant rats and the NW non-resistant rats. Similar to with the pattern observed for the *Vkorc1* region, in terms of the LD signals, we observed that: NW(susceptible) > NW(all) > NW(resistant) > LH (Figure 3.9D and E). This finding implies that selection has acted on this region; again, and it suggests that susceptible samples in a population that underwent selection would provide useful signals for detecting adaptive variations. The LD patterns were used as complementary source of information used to detect candidate genes. For example, though no candidate genes were identified on chromosome 1: 240-259 Mb region, there are LD block signals; this region covers some cytochrome P450 genes: *Cyp2c22*, *Cyp2c23* and *Cyp2c11*. *Cyp2c11* (cytochrome P450, subfamily 2, polypeptide 11) is the most possible ortholog of the human biomarker gene *CYP2C9* (Appendix 4). And *Cyp2c22* (cytochrome P450, family 2, subfamily c, polypeptide 22) is closely clustered with the *CYP2C9* group in the phylogenetic tree (Thomas 2007).

3.4. Discussion

3.4.1. GWAS – Bayes factors measure association strength

With the null hypothesis H_0 of no association, there are several advantages of the imputation based Bayesian statistical method. First, by computing the Bayes Factor (the ratio of probability under H_1 and probability under H_0) for each SNP, Bayesian formula "posterior odds = prior odds × BF" naturally integrated information across SNPs (Servin and Stephens 2007); thus yield more reliable results. Moreover, the Bayes Factor, as the likelihood ratio between H_1 and H_0, could assess the association strength. This feature makes the Bayes Factor as a good measure of association, which is used as gene score in the Network-guided GWAS analyses.

Second, P-values obtained from the standard association test are affected by factors such as the size of tests and MAF (minor allele frequency) (Stephens and Balding 2009). But removing low-MAF SNPs would introduce the risk of excluding meaningful associations. Besides, the methods to account for multiple testing, such as Bonferroni correction, assume that no true associations exist in the genome, which is not a reasonable in practice because as in our case study we 'knew' *a priori* of at least one significant association and numerous others due to genetic hitchhiking. Thus, P-values have limited value to detect true associations. Using Bayes Factor, the problem raised in multiple testing is alleviated (Stephens and Balding 2009).

Third, taking advantage of the linkage disequilibrium (LD) among nearby SNPs, unmeasured genotypes could be estimated ("imputed"). And the imputation process

naturally controlled for the confounding factor of LD among nearby SNPs, and thus is expected to have increased our power to detect associations (Servin and Stephens 2007).

We applied two genetic models to compute Bayes Factors: BF1 with the additive model (σ_d (dominance effect) = σ_a (additive effect) / 4 as suggested (Servin and Stephens 2007; Manna, Martin, and Lenormand 2011) and BF2 with the dominance model ($\sigma_d = \sigma_a$). As reported in the study of the *Vkorc1* sweep region (Chapter 2), *Vkorc1* is under over-dominance selection (c.f. Chapter 2). During the Bayesian association analysis, different selection models yielded different Bayes Factors; for *Vkorc1* the over-dominance model had the strongest signal. Thus, here for the genome scan, we considered the dominance model in addition to the general additive model.

3.4.2. NetGWAS facilitates candidate identification

There are several measures that can be used to quantify the role of genes in the network: 1) Degree centrality simply counts the number of neighbor genes connected to one node; 2) Closeness centrality is the inverse of a gene's farness, which is the sum of its shortest path distance to all other nodes; 3) Betweenness centrality is the probability of one gene occurring on a randomly chosen shortest path between two randomly chosen nodes; 4) Eigenvector centrality is a measure of a gene's influence by considering both the quantity and quality of its neighbors because not all connection are equal in the context of a phenotype (Davis et al. 2010). Google's PageRank algorithm computes the eigenvector as page ranks based on the number of pages linked to it and the ranks of these (Page et al. 1999).

Previous studies have successfully modified and applied the PageRank algorithm to Gene Ontology network analysis and transcriptional network analysis (Morrison et al. 2005; Winter et al. 2012). Here we adapted it to an analysis of a gene-gene interaction network to expand the traditional search for candidate genes as implemented by GWAS. This algorithm naturally integrates association measures as gene features that are allowed to spread in the network. The gene ranks are then computed as eigenvector by resolving a linear system of network topology and gene scores. Thus, computed gene ranks are not only dependent on the network topology but also are determined by gene scores (here: association measures). There is another network-based method called Network Interface Miner for Multigenic Interactions (NIMMI), which first computes gene ranks solely based on network topology, and then using the Liptak-Stouffer formula to combine gene ranks with z-scores converted from association P-values (Akula et al. 2011). This combination, however, may not be a true integration of the trait/phenotype association into genetic networks.

Google PageRank set the damping factor as 0.85, which is not suitable for much more sparse biological networks. Previous studies have used damping factor $d = 0.3\sim0.95$ (Morrison et al. 2005; Winter et al. 2012). We tried $d = 0.3$, evaluated as a good choice in transcriptional network (Winter et al. 2012). We also computed a dynamic damping factor for each node as suggested (Fu, Lin, and Tsai 2006) and used the average value *ddf* as damping factor for the whole network. The *ddf* for Network1 is 0.03 and *ddf* for Network2 is 0.02, which again reflected the sparse interaction in GGI network. We combined the rank results from both $d = 0.3$ and ddf to increase the reliability.

3.4.2.1. Limitation of the gene ranking

It is important to be aware of some limitations of network-based ranking approach. First, there are biases of the genetic interaction information (Barabasi, Gulbahce, and Loscalzo 2011). Some widely studies genes have more connections, whereas some new identified genes have less or none. Second, some genes will be given high rank due to network topology rather than their association with trait. We justified this situation by correcting the gene ranks using random shuffled gene score. Third, if SNPs are located at intergenic region (> 200 kb to the nearest gene), these SNPs cannot be mapped on genes, thus the following network analyses cannot be conducted. Therefore, we kept the candidate SNPs selected from the traditional GWAS approach that cannot be assigned to candidate genes for future reference (data not shown). Nevertheless, network-based gene ranking not only prioritized genes with strong association signals, and also expanded our search for candidate genes in case that some genes are missing initially because of low SNP density. As the improvement of function annotation and protein-protein interaction identification, the quality of genetic network information will increase, and thus the power of network-based approach.

3.4.3. Population genomic analysis supports 6-8 candidate regions

3.4.3.1. Measures of haplotype structure require higher SNP density

Population genomics signals are valuable for detecting adaptive variations among populations under different environments (Hernandez et al. 2011). Composite-likelihood ratio test based on genetic polymorphism has been commonly used to look for complete

sweeps and haplotype-based methods have been developed to discover ongoing sweeps (Kim and Stephan 2002; Sabeti 2007). The two measures of haplotype structure, iHS and XP-EHH, have been successfully used in previous studies (Voight et al. 2006; Sabeti 2007). In our situation of rat 10k array, however, the resolution of these haplotype-based measures are limited because of the low SNP density. Along the genome, we only obtained 501 iHS scores for the resistant population NW and 129 XP-EHH scores between NW and the non-resistant population LH. We assigned the iHS and XP-EHH to the candidate regions (Appendix 4) as supporting evidence rather than signals for detection.

3.4.3.2. Linkage disequilibrium (LD) signals selective sweep

LD blocks are important features for detecting adaptive variations (c.f. Figure 2.2A) (Kohn, Pelz, and Wayne 2000; Kim and Nielsen 2004). Here by splitting NW population that was under warfarin selection into two subgroups with different phenotypes, we surprisingly found that LD signals in susceptible samples are stronger than in resistant samples. By simulation and true data analyses, we showed that LD of susceptible rats could provide strong signals of selection. In general, this recognition reminds us to consider susceptible samples in experimental design and analyses.

3.4.3.3. Conclusions

With the aim to understand the genetic architecture of adaptation towards anticoagulant drug warfarin, we performed genomic scan for genotype-phenotype association signals. Based on traditional GWAS approach, we listed 82 SNPs with

relatively high association strength; two regions (chromosome 3: 157-158 Mb and chromosome 5: 54-56 Mb) with ≥ 3 candidate SNPs were detected. Modified Google's PageRank algorithm combines network information and gene-trait association to identify candidate genes by computing gene ranks. Among 87 identified candidate genes, 48 genes were clustered in 10 regions. Population genomic analysis found supports for 10 candidate regions based on haplotype structure measure and 6 blocks of linkage disequilibrium. Considering the gene-gene interaction and chromosomal proximity, we found that most candidate genes are directly or indirectly connected to vitamin K pathway, forming a 'warfarin resistance module' centered about the resistance gene *Vkorc1*. This picture depicts the genetic architecture of multiple genes involved in warfarin related pathways, which could be viewed as a single selective event having resulted in a multigenic response, but with a focus on sets of interacting genes.

Chapter 4

Polygenic adaptation at the gene expression level to warfarin selection in the Norway rat

Abstract

Genome-wide expression profiling potentially can reveals signals of adaptation that are manifest at the level of gene expression. In a wild-derived strain of Norway rats, we study the warfarin related expression changes across the genome. With the experimental design that contrasts resistant and susceptible phenotypes, each of these warfarin induced and non-inducted and represented by both sexes, here we built weighted co-expression networks and identified 591 candidate genes that displayed significant expression variations related to warfarin resistance phenotype (as determined by blood clotting response tests, BCR), warfarin resistance genotype (Y139C mutation in *Vkorc1* as determined by DNA analyses) and warfarin exposure as measured by injections of sublethal doses of the anticoagulant. These candidate genes formed 21 clusters. By comparing with previously identified candidate genes from a NetGWAS (Chapter 3), we identified 7 regions of clustered genes with support from both the NetGWAS analysis of SNP array data and the co-expression analysis conducted on RNA microarray data. Interestingly, although the mutation in the resistance gene *Vkorc1* is an amino-acid change with no known effects on gene expression, the region where the *Vkorc1* gene maps on chromosome 1 was identified as one of the 21 clusters. This is explained by high levels of *cis*-regulatory genetic polymorphisms underlying gene expression variation in wild populations of Norway rats, and their genetic hitchhiking with the Y139C mutation. This finding shows the promise to identify other regions with genes directly or indirectly affected by warfarin based on the genetic hitchhiking of SNPs that cause gene expression changes. Here we detected such *cis*-eQTLs by associating results of a genome wide single nucleotide polymorphism (SNP) analysis with an analysis of gene expression

profiles to obtain a list of genes that appear to form parts of the complex genetic architecture of warfarin resistance in the Norway rat.

4.1. Introduction

Gene expression is often considered as an important "intermediate level" connecting genotype with phenotype. Recent genome wide association studies of human complex diseases revealed many trait associated non-coding variations, which might be functional in gene regulation (Gilad, Rifkin, and Pritchard 2008). Gene expression analyses are applied to a wide range of medical research directions and thereby have greatly facilitated the identification of biomarkers and the elucidation of the genetic mutations causing diseases (Schadt et al. 2005).

From the evolutionary biological perspective, gene expression regulation is an important mode of adaptation. For example, polymorphisms in 5' upstream sites of the lactase gene regulate gene expression and contribute to the ability of human adults to digest milk; and in another system regulatory variation at the blood coagulation factor VII was reported for human populations under selection (Hahn et al. 2004; Tishkoff et al. 2007; Kudaravalli et al. 2009). Analyses of gene expression profiles have made progresses in understanding the genetic basis of adaptation in human populations, especially when recently associated with genetic variation data (Morley et al. 2004; Dixon et al. 2007; Stranger et al. 2007; Gilad, Rifkin, and Pritchard 2008; Kudaravalli et al. 2009). In both biomedical research and evolutionary biology a key question in need of

empirical data refers to the number and type of genes (genetic architecture) involved in the expression of phenotypic trait; be it a disease or an adaptive trait.

Previous analyses of differentially expressed genes have identified numerous diseases-associated candidate genes and biomarkers, individually and independently (Jafari and Azuaje 2006; Maertzdorf et al. 2011). However, working interactively rather than in isolation, genes act in concert to constitute expression and regulatory networks, which are translated (in response to cellular and extracellular cues) into phenotypes (Wittkopp 2007). Therefore, a network-based analysis approach appears to be a natural choice for the analysis of gene expression data. In fact, a co-expression network with genes as nodes and expression similarities as edges is inherent in microarray data. In such networks the highly connected clusters of genes with similar expression profiles stand out and thus can be identified individually, and analyzed as modules, in terms of their association with a traits value (Langfelder and Horvath 2008).

Second, incorporating known genetic interaction information can help further refinements of candidate gene lists. As in Chapter 3 here a modified version of Google's PageRank algorithm is explored for its use in gene ranking; i.e. by considering each gene's relevance to the traits and based on its interacting neighbors in network (Morrison et al. 2005; Winter et al. 2012). As exhibited during Google's searches PageRank returns user desired pages by ranking them based on the ranks of other pages linked to them (Page et al. 1999). With similar logic applied to genomic analyses the genes that are be given a high rank are those whose neighbors in a gene-gene interaction network also have high ranks. Here, such ranks are initially determined by the gene expression-trait

association that is used as the relevance measure, and then such ranks are computed iteratively by considering all the genes and their network positions and trait associations. Using this algorithm genes highly relevant to the trait under study are prioritized. The biological network is constructed based upon any type of genetic interaction information, such as protein-protein interaction, gene-chemical interactions, or transcriptional co-expression network.

Here we examined gene expression data collected for a warfarin resistant Norway rat (*Rattus norvegicus*) strain and analyzed the data for correlations with the resistance phenotypes and genotypes and exposure to the anticoagulant rodenticide warfarin (Hans-Joachim, Detlef, and Gerhard 1995; Kohn, Pelz, and Wayne 2000).

Warfarin caused fatal hemorrhage in susceptible rats as they fail to recycle vitamin K from vitamin K 2,3-epoxide. Specifically, warfarin inhibits the activity of the vitamin K 2,3-epoxide reductase complex (VKOR), which is encoded by *Vkorc1* (Li et al. 2004; Rost et al. 2004). It thus impairs the vitamin K cycle as well as the following post-translational modification process of multiple vitamin K dependent proteins needed for proper levels of activated blood-coagulation factors (Presnell and Stafford 2002; Stafford 2005).

The use of warfarin since 1950s has imposed intense selective pressure on rat populations (Endler 1985; Gillespie 1991; Kohn, Pelz, and Wayne 2000). However, resistance to warfarin in rat populations has been reported from numerous locations of the world (Boyle 1960; Lund 1964; Jackson and Kaukeinen 1972; population and agriculture 1986). The resistance phenotype can be measured by blood clotting response (BCR) tests

(Kohn, Pelz, and Wayne 2000). Genetically, the resistance trait in European rats is primarily determined by the Y139C mutation on the third exon of *Vkorc1* (Kohn, Pelz, and Wayne 2003; Rost et al. 2009).

In humans, warfarin is widely used as an antithrombotic drugs in the prophylaxis and treatment of recurrent stroke, deep vein thrombosis and heart valve prosthesis (Wadelius et al. 2007). However the proper dosing of the drug is a troublesome affair (Kamali 2006). Two genes, *VKORC1* and *CYP2C9* (and enzyme that metabolizes the S-enantiomer of warfarin) are approved by the Federal Drug Administration (FDA) as genetic biomarkers. The combined analysis in human subjects can explain ~30% and ~10% of warfarin dosage variance, respectively (Rost et al. 2004; Takeuchi et al. 2009). Recently, *CYP4F2,* which metabolizes vitamin K_1, has been reported to influence warfarin dosing also, but to a smaller degree than the other two genes that are used as biomarkers in personalized medicine (Takeuchi et al. 2009). Examination of other candidate genes thought to be involved in warfarin dependent pathways failed to yield new candidate genes that merit further analyses in human study cohorts (Wadelius et al. 2007). Another genome-wide scan for genetic variants that might influence warfarin dosing yielded no further genes either (Cooper et al. 2008). Some genes such as *GGCX, EPHX1* have been reported with controversial association results from different studies in different populations (Wadelius et al. 2005; Pautas et al. 2009; Takeuchi et al. 2009). But so far only ~55 % of warfarin dose variance can be jointly explained by known genetic and non-genetic factors such as sex, weight, age, etc.; leaving another ~45% of the variance around proper warfarin drug dosing in humans unexplained (Cooper et al. 2008).

In this study, we aim to identify candidate genes that differ in their gene expression patterns (collected for the liver tissue) with regard to warfarin exposure, warfarin resistance genotype, and sex. First we built co-expression networks in 4 different comparisons considering resistance phenotype and warfarin treatment. By correlating expression profiles to warfarin-related traits in the co-expression network, we identified network modules and the candidate genes that form them. Further we combined expression-trait correlation measures with genetic network information (including data collected from published gene-gene interaction networks and a gene-chemical interaction network). To evaluate any functional relationships among genes we conducted functional analyses and pathway annotations. Finally, we compared the results of this study of gene expression analysis with our previous network-guided genomic variation analysis (c.f. Chapter 3) and detected a strongly supported set of genes that are part of eQTL (expression quantitative trait loci). As in Chapter 3, this thesis chapter 4 further supports the hypothesis that warfarin resistance has complex genetic underpinnings, with many of these underlying genetic factors being a target of natural selection.

4.2. Materials and Methods

4.2.1. Design of microarray study

This study investigates two parameters for their effects on gene expression in the liver of rats (Figure 4.1A). Wild-derived lab-reared rats were resistant to the anticoagulant warfarin owing to a non-synonymous mutation in the *Vkorc1* gene

(Y139C). The microarray experiment was designed for us to investigate the influence of warfarin treatment and warfarin-resistance phenotype on gene expression.

4.2.2. Rat strains

This study was conducted on rats sampled from an area in Northwestern Germany. Rats from this area carry a non-synonymous mutation in the *Vkorc1* gene from A to G, resulting in the tyrosine to cysteine (Y to C) amino-acid change at position 139 in exon 3. This amino-acid change results in a 47% reduction of the basal in vitro VKOR activity (Pelz et al. 2005; Rost et al. 2009). The *Vkorc1* genotypes are designated as G/G (homozygous mutant), G/A (heterozygous mutant) and A/A (wildtype). The strain examined is resistant to the anticoagulant warfarin if *Vkorc1* carry the genotype of G/G or G/A, i.e. the mutation is dominant with respect to warfarin resistance (but is of incomplete penetrance, c.f. Table 2.1).

Rats, named as NW rats, were trapped in the field and were kept in the laboratory for 12-17 months, or rats were born by wild-caught rats in the laboratory (age was recorded in Supplementary Table 1 of (Kohn, Price, and Pelz 2008)). Rats were kept in Macrolon® cages on shavings from saw mills with tap water and standard lab food (altromin® 1324) ad lib. The diet contained 3 mg/kg Vitamin K3. The Institutional Animal Care and Use Committee approved the study.

The warfarin resistant (R) and warfarin-susceptible (S) phenotypes were distinguished by using a blood clotting response test (BCR) that measured percent clotting activities (PCA) after injection of a diagnostic dose of warfarin (c.f. Rodenticide

Resistance Action Committee 2003, a reappraisal of blood clotting response tests for anticoagulant resistance and a proposal for a standardized BCR test methodology: Technical Monograph, Crop Life International). Male and female rats were induced with 7mg/kg or 8.5mg/kg warfarin, respectively, by intraperitoneal injection, which corresponds to 4 multiples of the ED50 in susceptible rats. This treatment unlikely explains the systematic differences among genotypes described in the results section as all genotypes were subjected to it. The presence of the A to G mutation at position 139 in exon 3 of the *Vkorc1* gene was confirmed by PCR.

4.2.3. Tissue processing

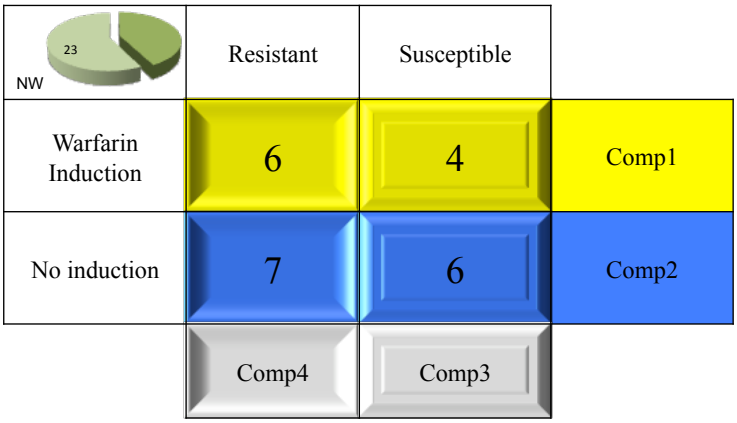
Liver is used as the tissue of choice because it is the organ where warfarin is metabolized (Thijssen 1995; Wallin et al. 2001), VKOR is located (Wallin et al. 2001; Rieder et al. 2005; Oldenburg et al. 2006), and important detoxification pathways are located (Runge-Morris et al. 1998; Raftogianis, Wood, and Weinshilboum 1999; Beckmann-Knopp et al. 2000; Tabrett and Coughtrie 2003; Wang et al. 2004; Wang, Zhao, and Dudoit 2006). Vitamin K cycle activity is traditionally measured from hepatic tissues and the VKOR is thought to reside in the liver microsomal membrane. Liver tissues are stored in RNALATER (Ambion Inc.) at -20C and shipped from Germany to Houston.

Total RNA was isolated from the liver of rats for array work. RNA quality control, labeling and hybridization of the RNA to the AFFYMETRIX rat genome ARRAY 230 2.0 were done by a commercial laboratory in Houston (SEQWRIGHT, http://www.seqwright.com/). The procedures to assay gene expression followed

recommendations by the supplier of arrays (http://www.affymetrix.com/). Quality control (QC) of whole RNA was done on an Agilent 2100 Bioanalyzer RNA LabChip.

4.2.4. Microarray data and pre-processing (Figure 4.1B)

A. Experiment design: grouped by phenotype and warfarin treatment



B. Co-expression network analysis

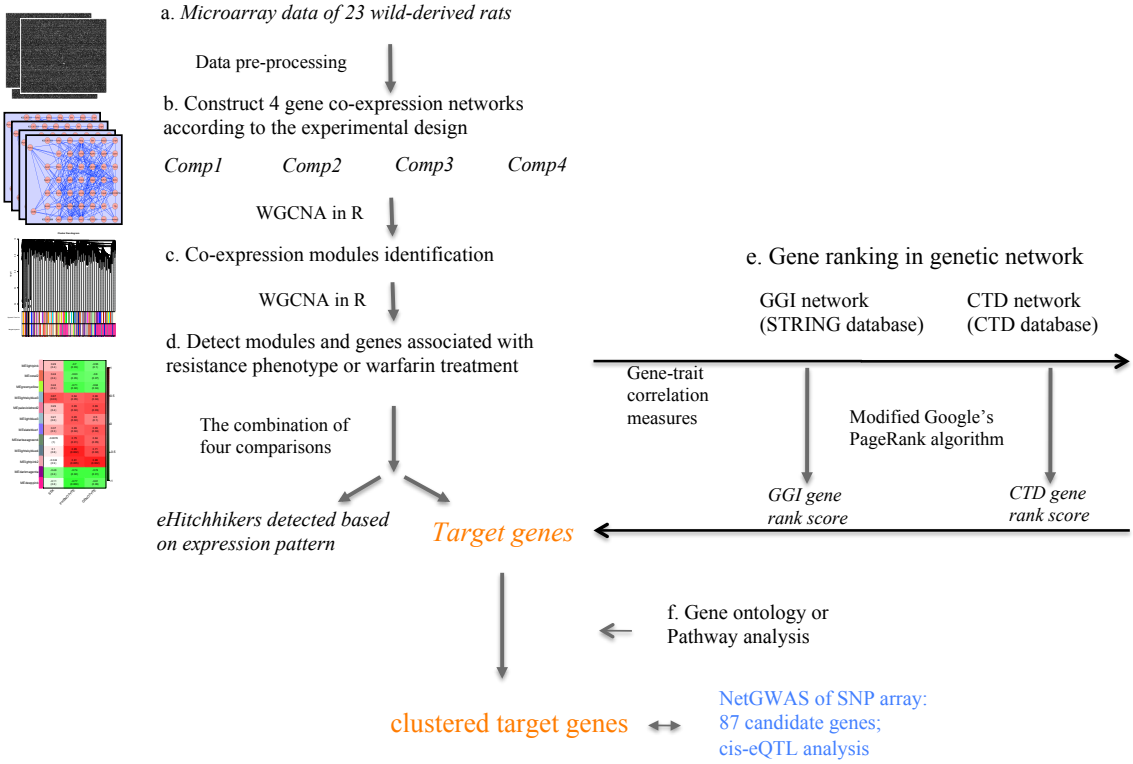


Figure 4.1 – The experiment design (A) and workflow (B) of network-guided expression analysis. Flowchart indicates the main steps of identifying candidate genes related to warfarin in rat co-expression networks.

Expression data of rat liver for 24 warfarin-resistant and warfarin-susceptible rats (14 males and 10 females) were obtained from Affymetrix arrays based on the platform of Affymetrix GeneChip Rat Genome 230 2.0. Corresponding rat genome information (NCBI build 4) was retrieved from UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables, accessed July 2010). Traits like sex, warfarin treatment and resistance phenotype were considered during each of the analyses (Table 4.1).

Table 4.1 – Microarray design and sample infromation.

Rat	ArrayID	Induction	Sex	Phenotype	Genotype	Comp1	Comp2	Comp3	Comp4
5783	B67	N	M	S	A A		CON	CON	
5991	B68	N	F	S	A A		CON	CON	
5989	B69	Y	M	S	A A	CON		TRE	
5994	B70	Y	F	S	A A	CON		TRE	
5738	C30	N	M	R	G G		TRE		CON
5791	C31	N	F	R	G G		TRE		CON
5766	C32	N	M	R	G A		TRE		CON
5909	C33	N	F	S	A A		CON	CON	
5744	C34	Y	M	R	G G	TRE			TRE
5750	C35	Y	F	R	G G	TRE			TRE
5761	C36	Y	M	S	A A	CON		TRE	
5801	C38	N	M	R	G A		TRE		CON
5658	C39	N	M	R	G G		TRE		CON
5758	C40	N	F	R	G G		TRE		CON
5768	C41	N	F	R	G A		TRE		CON
5907	C42	N	M	S	A A		CON	CON	
5784	C43	N	M	S	A A		CON	CON	
5910	C44	N	F	S	A A		CON	CON	
5751	C45	Y	M	R	G G	TRE			TRE
5745	C46	Y	M	R	G G	TRE			TRE
5760	C47	Y	F	R	G G	TRE			TRE
5678	C49	Y	M	R	G A	TRE			TRE
5731	C50	Y	M	S	A A	CON		TRE	

Induction (Y/N) refers to warfarin induction Yes or No;
Sex (M/F) refers to Male or Female;
Phenotype (R/S) refers to Resistance to warfarin or Susceptible to warfarin;
Genotype refers to the nonsynonymous mutation Y139C of *Vkorc1* gene.
Four comparisons (Comp1, Comp2, Comp3 and Comp4) are designed to detect modules and genes associated with phenotype and induction of warfarin. TRE-treatment, CON-control.

Data background correction, normalization, probe specific background correction and summarization were conducted using the Affy software version 1.18.2 (Gautier et al. 2004) and the gcRMA version 2.12.1 (Wu et al. 2004) package in R (Ihaka and Gentleman 1996). Probes were matched to genes using the Rat230_2 annotation file (release 32, 06/09/11) obtained from Affymetrix (http://www.affymetrix.com/). After quality control examination, we removed one outlier female array C37 (sample 5675), resulting in 23 arrays of 31,099 probes. PCA analysis verified that the biological variables (warfarin induction, sex, and genotype) were indeed dominating the variation in expression, but the extend to which the gene expression over the rat genome was affected was unexpected given that *Vkorc1* Y139C mutation is a coding mutation and apparently is the main genetic factor underlying resistance. Such analyses enable an exclusion of the effect of experimental issues, such as variation of expression results owing to sampling, extraction procedures, or even shipping of samples between laboratories.

To reduce false positive rates of gene identification we filtered out low quality data based on the Affymetrix detection calls provided alongside CEL files. First, for each probe set the detection call (Absent (A), Present (P) or Marginal (M)) was calculated using Wilcoxon signed rank test (mas5calls in affy package) to measure the significance level (P-values) of signal intensities of the perfect match probes compared to the signal

intensities obtained for the mismatch probes. As suggested (McClintick and Edenberg 2006), we filtered out probe sets with 'Absent' or 'Marginal' call in at least 25% of the samples treated with or without warfarin respectively. After data processing 15,531 probes remained.

Finally, we excluded the control probes with the prefix AFFX in probe labels. In sum, here we report on the analyses of 23 arrays and 15,480 probes each. Expression data are deposited to the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/).

4.2.5. Traditional analysis of gene expression using PADE

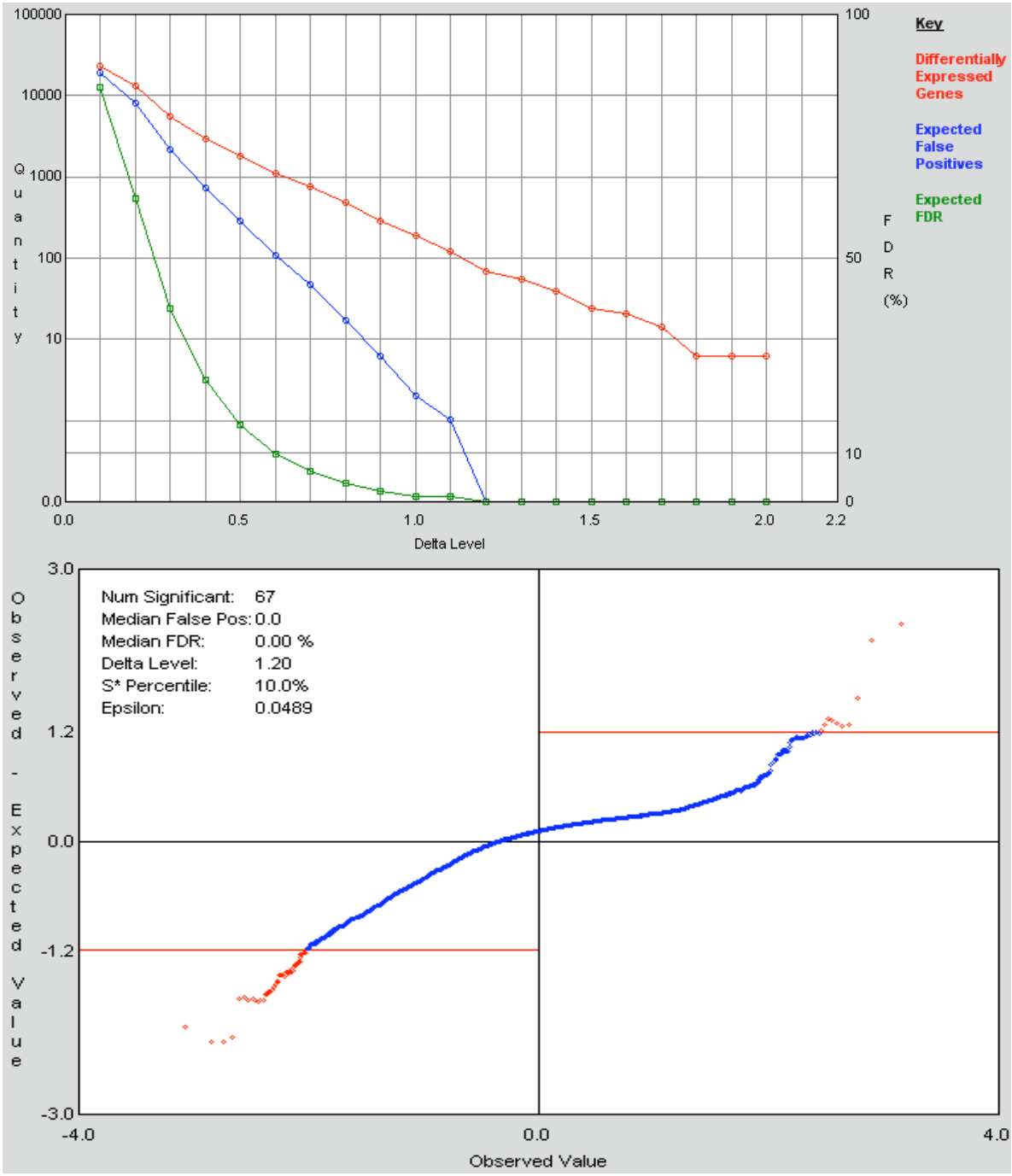
Basic statistical identification of differentially expressed genes used the AFFYMETRIX SOFTWARE MAS5 perfect match (PM) and mismatch (MM) probe analysis after application of REDI (REDUCTION OF INVARIANT PROBES) ANALYSIS. REDI was used to address issues with the reliability of expression analyses that were traced back to the sequence variation in the AFFYMETRIX probes on the array. The probes (not the entire probe set/transcript) are eliminated if they exceed a threshold level of noise, i.e. those that do not correspond to the expression level inferred from the majority of the other probes representing the transcript. We employed the algorithm by EXPRESSION.COM, DURHAM, NC, USA; http://www.expressionanalysis.com/) to identify and exclude poorly performing probes from analysis. In this two-step approach poorly performing perfect match (PM) probes were removed.

We used the feature TWO-GROUP COMPARISONS WITH PERMUTATION ANALYSIS FOR DIFFERENTIAL EXPRESSION (PADE, EXPRESSION.COM, DURHAM, NC, USA) to detect changes in expression between groups with different phenotype and warfarin treatment.

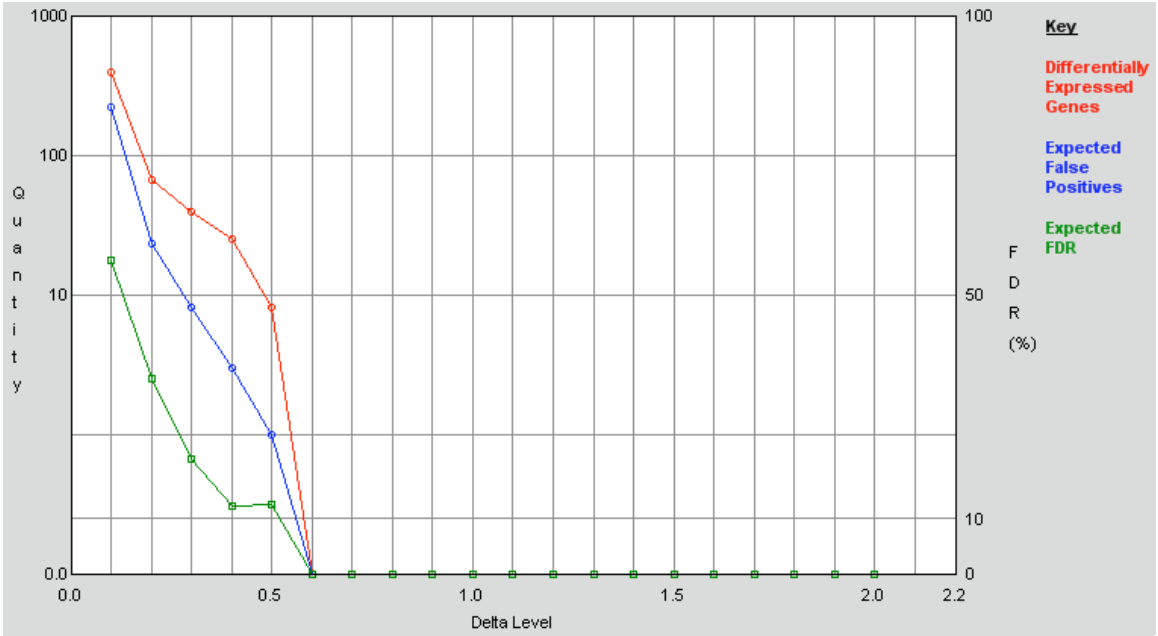
Permutation based analyses enable estimation of the probability of falsely detecting differential expression, i.e. False Discovery Rates (FDR). We estimate the number of genes that is affected by an experimental treatment using the FDR as a guide. The FDR should be interpreted as the fraction of expected false positives in a given set of genes, without making statements which genes are the false positives. However, PADE attempts to do just that by permutation analysis that generates thousands of smaller subsets of genes and isolating those that among thousands of such small sets of genes consistently stand out. In short, we will use the FDA predominantly to qualify our statements regarding the genome wide response to warfarin induction. FDR graphs (Figure 4.2) are used to obtain the number of significant genes given a particular FDR and provide a (delta) cutoff value used to isolate significant genes.

Comparisons considering warfarin treatment and resistance phenotype were analyzed in PADE. The differentially expressed genes identified from two comparisons using the corresponding delta cutoff (delta ≥ 1.2 in resistant vs. susceptible phenotype comparison; delta ≥ 0.4 in warfarin induction vs. non-induction comparison) were combined and used for future validation. We also selected 101 differentially expressed genes using P-value (≤ 0.05) from two comparisons (intersection) and compared these results with the co-expression network analysis results.

A. Comparison between resistant vs. susceptible rats.



B. Comparison between warfarin induction vs. noninduction.



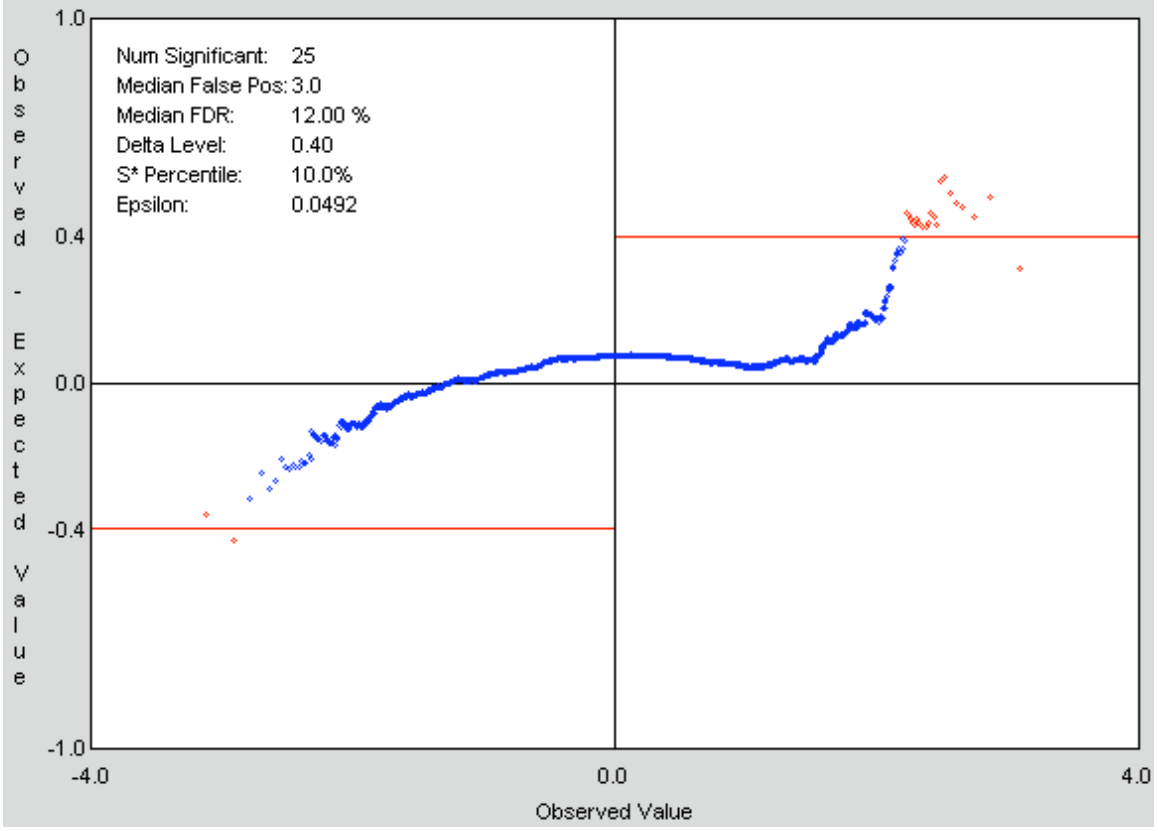


Figure 4.2 - False discovery rate plots suggest cutoff for differentially expressed genes. (A) Comparison between resistant vs. susceptible rats. (B) Comparison between warfarin induction (treated with warfarin) vs. non-induction.

4.2.6. Co-expression network and weighted correlation analysis

By constructing a co-expression network (Figure 4.1B) based on the rats microarray data using R package WGCNA (Langfelder and Horvath 2008), we aim to identify candidate genes responding to warfarin induction and resistant/susceptible phenotype. In co-expression network, each gene is a node and the edge between genes represents their expression similarity.

First, data cleaning: we removed genes with too many missing values with the default settings and verbose = 3.

Second, network construction: We built co-expression network by calculating the co-expression similarity $S_{ij} = | cor(g_i, g_j) |$ between gene pairs and adjacency (connection strength) matrix $a_{ij} = s^{\beta}_{ij}$ with β chosen as the lowest integer resulting in network with approximate scale-free topology (linear regression model fitting signed R^2 between degree distribution log(p(k)) and average degree $\log(k) \geq 0.9$ (in Comp1 $\beta = 9$ reaches plateau at 0.8). This criterion of soft threshold β was applied to ensure co-expression network property satisfying scale-free architecture (Ghazalpour et al. 2006). Contrary to unweighted network with absolute value 0 (no co-expression) and 1 (perfect co-expression) determined by “hard” threshold, here our weighted correlation with soft threshold preserves the continuous nature of the gene expression data.

Third, module detection: we used unsupervised hierarchical clustering method to identify clusters of densely interconnected genes, i.e. modules, with a minimum module size equals 30 probes (Langfelder and Horvath 2008; Ivliev, Rudneva, and Sergeeva 2010). For each model, the eigengene was identified as the first principal component of the expression matrix. Then modules with very similar expression profiles were merged based on the correlation between eigengenes.

Fourth, after identifying characteristic profile (eigengene) for each module, we look for modules significantly associated with external trait (resistance phenotype or warfarin induction) by calculating module significance (MS) based on the correlation of eigengene profiles and external traits. Specifically, two correlation measures were computed by Pearson correlation and Logistic regression respectively. The modules that significantly correlated with traits were shown in Figure 4.3. Meanwhile, for each probes, gene-trait association significance (GS) was also calculated as either the Pearson correlation coefficient or the pseudo r-squared of Logistic regression. These correlation measures of gene-trait association were also used as gene scores for the following gene ranking analyses. A fuzzy measure of membership of each gene in each module was also calculated. In addition, for each gene, we computed the overall network connectivity (K_{all}) as the sum of connection strengths with all other genes, i.e. node degree. To describe gene-module structure, we also calculated intramodular network connectivity (K_{in}) and outmodular connectivity (K_{out}) as well as the difference ($K_{dif} = K_{in} - K_{out}$) of each gene across modules. The connectivity measures for each module were averaged across genes within the module.

Following this process, we constructed 4 co-expression networks for following comparisons (Figure 4.1A):

Comp1: detect genes associated with resistant or susceptible phenotype when exposed to warfarin (10 samples of 15081 probes, softPower $\beta = 9$);

Comp2: detect genes show basal transcriptional difference between resistance and susceptible phenotype without warfarin induction (13 samples of 15125 probes, softPower $\beta = 6$);

Comp3: detect genes responding to warfarin administration in susceptible rats (10 samples of 15020 probes, softPower $\beta = 7$);

Comp4: detect genes responding to warfarin administration in resistant rats (13 samples of 15171 probes, softPower $\beta = 9$).

The slight differences of probes numbers among comparisons were due to the initial data cleaning of missing values. The combination of above four comparisons (Figure 4.1B) helps us identify candidate genes of warfarin, as well as their surrounding neighbors affected by the potential causal genes.

First, candidate genes were expected to show differential transcriptional patterns between resistant and susceptible rats (Comp1 or Comp2); and also these genes would respond to warfarin induction either in susceptible rats or in resistant rats (Comp3 or Comp2).

$$\text{Target} = (\text{Comp1} \cup \text{Comp2}) \cap (\text{Comp3} \cup \text{Comp4})$$

Equation 4.1 – The combination scheme for identifying candidate genes.

Genes affected by target genes because of their physical closeness to targets might also show different expression patterns between resistant and susceptible rats despite of warfarin induction; these genes, however, would not respond to warfarin administration in neither resistant nor susceptible rats. Similar to hitchhiked variants under selective sweep effect, these genes were called ehitchhikers here:

$$eHitchhiker = (Comp1 \cup Comp2) \cap \overline{Comp3} \cap \overline{Comp4}$$

Equation 4.2 – The combination scheme for identifying ehitchhikers.

We first selected candidate genes from each comparison with the criteria that the gene itself and the module it belongs to are significantly associated with traits (P-value ≤ 0.05). The modules significantly correlated with traits in each comparison were shown in Figure 4.3. Then we applied above combination schema (Equation 4.1 and Equation 4.2) to all the comparisons and obtained a list of candidate genes (Appendix 6) and ehitchhikers (data not shown). Note that here the candidate genes and ehitchhikers are independently defined based on their expression correlation with phenotype.

Several regions with clustered candidate genes were noticed, which might be the signals of highly related target genes in the region. Thus we defined regions of clustered gene as more than 5 genes located in a region with neighbor genes' distance smaller than 2mb (listed in Table 4.2 and Appendix 6). We plotted target genes along the genome (Figure 4.6), from which signals of candidate targets' clusters (dense black bars) could be observed.

4.2.7. Gene ranking using known genetic interaction information

Above analyses were based on co-expression networks. In this section, we combined prior knowledge of network information with the expression-trait correlation measure to refine the candidate list. As introduced in Chapter 3, the modified Google's PageRank algorithm was applied to two types of networks for gene ranking.

4.2.7.1. Gene-gene interaction (GGI) network

In GGI network, each gene is a node, and the edge between genes represents their protein-protein interaction or other type of interactions. We downloaded the interaction information from STRING (functional protein association networks) database (http://string.embl.de/, accessed July 2012) (Jensen et al. 2009), which include protein links from diverse channels: neighborhood, gene fusion, occurrence, co-expression, experiments, database and text mining. We only retrieved the protein interactions experimentally derived or from databases of physical interactions or biological pathways or based on text-mining with the threshold of confidence score ≥ 150 applying to each channel. The remaining information contains 15376 proteins and 344481 interactions. Then we converted proteins into corresponding genes and matched them with the genes from microarray data in each comparison. Within each GGI network for each comparison (Table 4.4), we applied the modified PageRank algorithm (c.f. Chapter 3: NetGWAS) to compute the gene ranks.

4.2.7.2. CTD (Comparative Toxicogenomics Database) network

We built CTD networks with warfarin related gene-gene interactions from CTD database (http://ctdbase.org, accessed Aug 2012). First, we downloaded the chemical vocabulary, 34 chemicals of them are warfarin related or the upper categories in the hierarchical classification system. Then, we found 46 genes interacting with above 34 chemicals from the downloaded rat chemical-gene interaction data. Also from the chemical-gene interaction data, we generated gene-gene interactions if genes are interacting with same chemicals. The interaction strength between two genes equals to their interacting chemicals. We kept the gene-gene interactions with one gene name on the list of 46 warfarin related genes, resulting in 69760 interactions. After that, we matched the genes available from microarray data in each comparison and built corresponding CTD network (Table 4.4).

Using the modified PageRank algorithm, we obtained gene rank files from both GGI and CTD network for each comparison. Then we assigned the gene rank scores to the selected candidate genes (Appendix 6). Then for each gene, according to Equation 4.1, we selected the higher rank score between Comp1 and Comp2, also the higher rank score between Comp3 and Comp4, then averaged them to get the final rankscore for candidate target. The rank scores from GGI and CTD network were listed besides candidate genes as reference. The 21 clusters of candidate genes were shown in Table 4.2 with average rank scores for each cluster.

4.2.8. Candidate genes for Function/Pathway evaluation

To evaluate the candidate genes' function, we performed GO (Gene-Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis in DAVID (http://david.abcc.ncifcrf.gov, accessed Aug 2012). We submitted a background gene list with all genes available from microarray data and a candidate gene list with 163 selected genes. The GO Fat category was chosen to avoid the overshadowing of broadest terms on more specific terms. KEGG and protein domain information (InterPro, PIR, SMART) were also included. 149 functional categories (of the 157 DAVID IDs corresponding to candidate targets) from GO, KEGG or Protein Domains databases were grouped into 45 function clusters. Their detailed information and corresponding genes were listed in Appendix 7. After removing the categories with the enrichment P-values > 0.05, we obtained 19 function clusters composed of 80 categories. We depicted the enrichment of genes in each functional cluster with the highest enrichment score of the functional category in that cluster for each gene (Figure 4.6).

4.2.9. *cis*-eQTL (quantitative trait loci) analysis

As suggested by Kudaravalli et al. (Kudaravalli et al. 2009), we tried to identify SNPs (single nucleotide polymorphisms) that are associated with gene expression profiles. *cis*-eQTL were defined as SNPs that located on a gene with boundaries of ± 100kb to transcription start and end, which significantly associate with expression profiles. Though without direct evidence of true *cis*-acting regulation, these SNPs were refer to as *cis*-eQTL since they might regulate expression levels of their corresponding genes (Kudaravalli et al. 2009).

Genotype data of 10847 SNPs were obtained based on Rat 10k array for the same 24 NW wild-derived rat samples (c.f. Chapter 3). 14 samples had both expression and SNP array data. Genotypes of SNP were codes into 0, 1, 2. The association between genotype and expression profile were tested using standard linear regression. Multiple SNPs could be mapped to one gene with multiple probes; in this case, the association of each SNP-probe pair was calculated. We consider SNPs with P-value ≤ 0.05 as *cis*-eQTLs and also selected a smaller set of SNPs with significance level threshold $\leq 10^{-4}$ as suggested (Kudaravalli et al. 2009).

With the identified *cis*-eQTL, we tested whether they were enriched in the candidate genes using hypergeometric test in R. Here the candidate gene list we tested includes the 591 candidate genes based on expression network analysis (Appendix 6), 101 or 55 differentially expressed genes using traditional expression analysis (data not shown), and 87 candidate genes based on SNP array analysis (c.f. Chapter 3). The significance levels were recorded in Table 4.5.

4.3. Results

With the goal of identifying genes related to warfarin, we designed four comparisons based on 23 rat (*Rattus norvegicus*) samples (Figure 4.1) and collected their RNA expression data on Affymetrix microarrays. What we label as Comp1 and Comp2 are the comparisons aimed to detect genes differentially expressed between resistant and susceptible rats and with or without warfarin induction, respectively. On the other hand,

Comp3 and Comp4 intended to identify genes that are differentially expressed with or without warfarin induction in either susceptible rats or resistant rats, respectively.

Co-expression networks of different comparisons were then built to detect candidate genes that are related to warfarin in terms of *Vkorc1* genotype, warfarin resistance phenotypes as measured by the BCR method, and sex. We incorporated gene-gene interaction and gene-chemical interaction information to facilitate the candidate identification using a modified Google's PageRank algorithm. The combination of co-expression network and prior knowledge on genetic network location and interactions of genes enabled us to prioritize candidate genes whose expression changes showed associations with the experimental variables studied here.

4.3.1. A co-expression network identifies ~600 candidate genes

The co-expression networks were built based on rat microarray data. Nodes are genes, and edges between them are the transformed expression similarity: connection strength = | similarity |$^{\beta}$ with β as the chosen soft threshold (see methods and materials). The connection strength of the network preserves the continuous nature of co-expression data, and the results of network analyses are robust to the choice of β (Ghazalpour et al. 2006). We built four co-expression networks according to the designed comparisons and intended to combine the results from them to detect candidate genes (Figure 4.1).

4.3.1.1. Modules were detected in the co-expression network

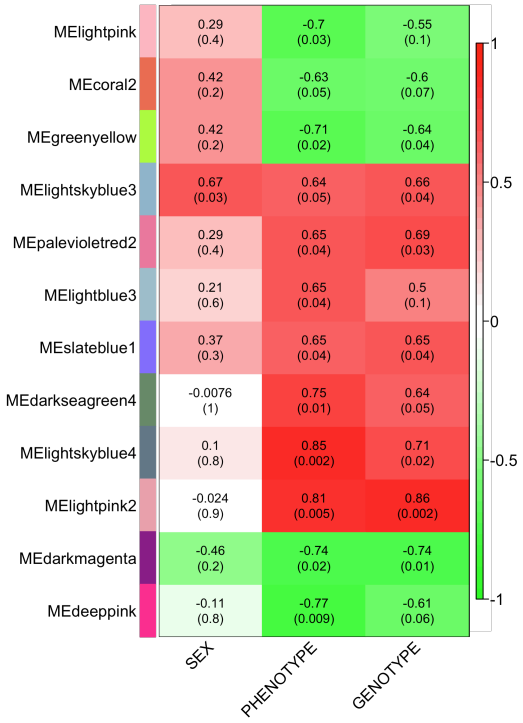
Co-expressed genes tend to involve in similar functions/pathways (van Noort, Snel, and Huynen 2003). Co-expression networks could be clustered into modules, which

respond to environmental change as a whole in terms of expression profiles. Thus, detecting modules and correlating them with external traits would facilitate the identification of candidate genes, whose expression changes might be related to traits. For each comparison, we identified modules with highly connected genes using hierarchical clustering. Modules were represented by different colors, and their properties were recorded but not shown here. For example, in Comp1 (resistant vs. susceptible rats with warfarin induction), module sizes ranged from 30 to 2,659; with an average size of 198 probes. Their overall connectivity (K_{all}, node degrees) ranged from 37 to 341 nodes across different modules, which is lower than the average $K_{all} \sim 95$ seen for the total network. We calculated two other measures of network cluster structure: intramodular connectivity (K_{in}) and outmodular connectivity (K_{out}). In Comp1, the average K_{in} and K_{out} across modules are 20 and 75, respectively.

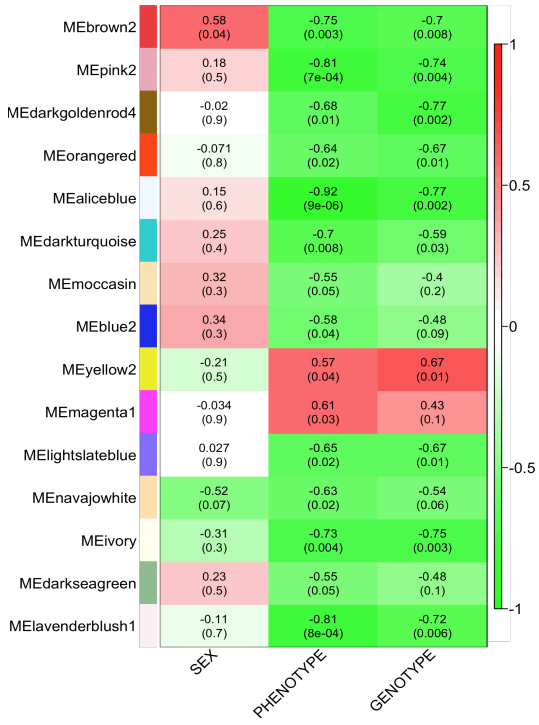
4.3.1.2. Modules significantly associated with traits

To assess the trait relevance of each module, we calculated the module significance (MS), which measured the correlation between traits and expression profiles of each gene in each module (see Methods and Materials). Figure 4.3 showed the modules that significantly correlate with traits in the four comparisons Comp1-4. In Comp1, the module most relevant to resistance was called lightskyblue4 (MS=0.85, P-value=0.002, with K_{in}=65.9, K_{out}=189.5, module size=840). The known resistance gene *Vkorc1* was located in the module called lightpink (MS=-0.7, P-value=0.03, with K_{in}=56.2, K_{out}=122.2, module size=645).

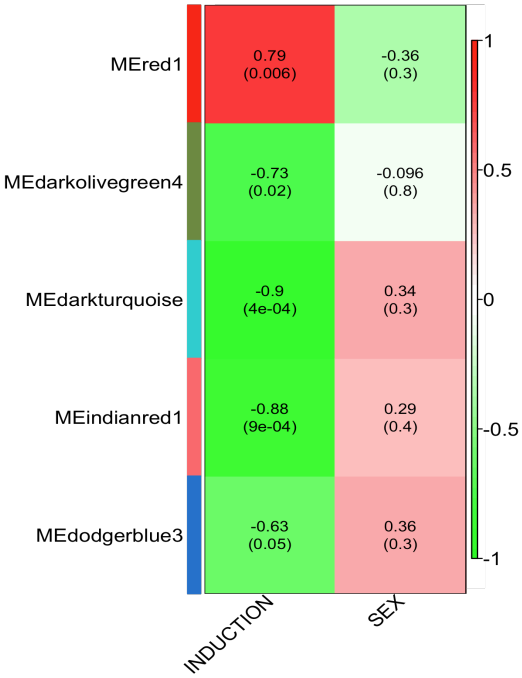
A. Comp1 (R vs. S rats with induction)



B. Comp2 (R vs. S rats without induction)



C. Comp3 (warfarin treatment in S rats)



D. Comp4 (warfarin treatment in R rats)

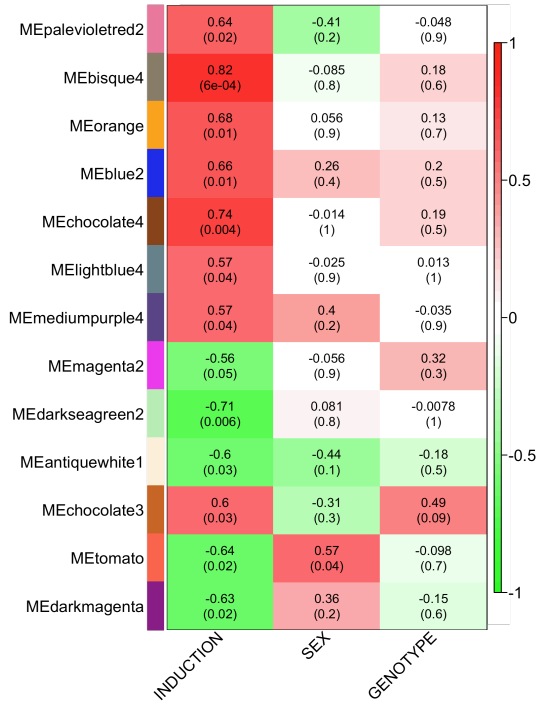


Figure 4.3 – Modules correlated with warfarin treatment and phenotype in the coexpression network. Modules (named by colors) are highly connected gene clustered identified in co-expression network in each comparison. Correlation coefficients of trait-module correlation (and the P-values) are shown.

Overall, there were 12 modules correlated with resistance phenotype with warfarin induction; while without warfarin induction, there were 15 modules correlated with the warfarin resistance phenotype in Comp2. In Comp3, only 5 modules significantly correlated with warfarin treatment in susceptible rats; whereas in Comp4, 13 modules were associated with warfarin treatment in resistant rats.

4.3.1.3. About 600 genes emerged from co-expression network analyses

Previous studies suggested anticoagulant rodenticide resistance might be attributed to expression changes of a few genes, such as those that belong the cytochrome P450 family (Markussen et al. 2008b). With the goal of detecting genes whose expression changes are related to the experimental variables considered here, we expected that target genes not only associate with resistance phenotype (Comp1 or Comp2) but also to warfarin induction (Comp3 or Comp4) (Equation 4.1). Thus, we first tried to select genes that are differentially expressed between resistant and susceptible rats (Comp1 $\cup$ Comp2); then we chose genes that are differentially expressed between the groups of rats that underwent warfarin induction and those that did not (Comp3 $\cup$ Comp4). After that, we intersected results to obtain a list of 719 probes, corresponding to 591 candidate genes (118 probes could not be matched to genes (Appendix 6). We expect to have false positive associations in this dataset.

Genes that map in the chromosomal neighborhood of genes that directly interact with warfarin induction might exhibit similar expression profiles as genes that are associated with the warfarin resistance genotype/phenotype (Comp1 $\cup$ Comp2) because of the hitchhiking effect. However, such genes, called ehitchhikers, would not respond to warfarin treatment ($\overline{Comp3} \cap \overline{Comp4}$). With this logic in mind (Equation 4.2) we identified 1,328 candidate ehitchhikers that were not directly related to the experimental variables genotype, phenotype, and induction; sex being non relevant in this context.

The candidate genes were selected not only based on their gene-trait association significances (GS), but they also must be located within the interaction network modules (MS) that were found to correlate with the experimental variables genotype, phenotype, and warfarin induction (Appendix 6). The known resistance gene *Vkorc1* did not emerge as significant data point (Comp1: lightpink module, P-value=0.486; Comp2: coral3 module, P-value=0.865; Comp3: tan2 module, P-value=0.064; Comp4: pink4 module, P-value=0.557). This is not surprising since the known Y139C variation confers resistance in form of a non-synonymous protein coding mutation (Rost et al. 2009). On the other hand, a suspected ehitchhiker, *Sult1a1* (unpublished data) with polymorphisms in its 5' upstream region was on the ehitchhiker list (Comp1: lightskyblue4 module, P-value=0.025; not significant in Comp3 and Comp4).

Table 4.2 – Candidate regions of clustered target genes in the rat genome.

Regions	Chr	Position (Mb)	GGIRank	CTDRank	Gene no.	Start & End Genes	Candidates (Chapter 3)
1	1	82 - 86	4	7	10	*Cyp2t1; Gpi*	
2	1	185 - 187	1	4	8	*Tufm; Pycard*	*Vkorc1*
3	1	201 - 213	8	1	22	*Psmd13; Prpf19*	

4	1	247 - 252	11	17	6	*Slc25a28; Taf5*	
5	2	178 - 182	7	8	6	*Rps3a; Snapap*	*Bglap*
6	3	154 - 158	13	17	6	*Fitm2; Ddx27*	*154 - 158*
7	5	136 – 140	4	8	6	*Mast2; Ppcs*	
8	7	8 – 13	8	17	13	*Zfp347; Cyp4f6*	*Cyp4f1*
9	7	67 – 69	20	20	5	*Cdk4; Hrsp12*	
10	7	113 – 116	2	2	12	*Ly6e; Cdc42ep1*	
11	7	137 – 140	15	2	5	*Tuba1b; RGD1359310*	
12	8	46 – 48	15	5	6	*Oaf; Sidt2*	
13	8	112 – 115	8	21	6	*Rassf1; Lrrfip2*	
14	10	70 – 73	21	5	5	*Tmem132e; Usp32*	
15	10	106 – 110	6	13	10	*Galk1; Sectm1b*	
16	11	82 – 83	2	13	6	*Chrd; Klhl24*	
17	12	33 – 37	13	13	5	*Setd8; Plbd2*	*Clip1*
18	14	84 – 87	15	8	7	*Tcn2; Ccm2*	
19	16	17 – 20	11	8	8	*Mbl1; Atp14a1*	
20	19	49 – 53	19	16	6	*Mlycd; Tcf25*	*Mlycd*
21	20	1 – 3	15	8	5	*Znrd1; RT1-CE5*	*RT1-S3*

Chr: chromosome.

GGIRank: the rank based on average gene rank score from GGI network for each gene region.
CTDRank: the rank based on average gene rank score from CTD network for each gene region.
Start & End Genes: the start and end genes within each candidate region.

Candidate genesfromGWAS: candidate genes detected in rat SNP array I (GWAS) located in the same candidate regions identified here from microarray data.

From the list of candidate genes (Appendix 6), we noticed that 8 candidate genes clustered (Figure 4.6A, bold red in Appendix 6) at the region of 185 -187 Mb on chromosome 1, which should be under the selective sweep effect of resistance gene *Vkorc1* (Kohn, Pelz, and Wayne 2000)(c.f. Chapter 2). We noted that although the resistance variant Y139C in *Vkorc1* is a non-synonymous mutation with no consequences on the expression profile of *Vkorc1*, the selective sweep associated with the warfarin selection on *Vkorc1* generated a broad genomic window in which numerous gene expression differences were assoiated with the experimental variables. Thus, we predicted that many genomic regions with spatially clustered gene expression differences

might indicate the location of either an adaptive gene expression change or the presence of another type of selected mutation (e.g. non-synonymous) within this region.

Thus we searched the data for regions of clustered candidate genes along the genome (Table 4.2; for gene information c.f. Appendix 6). On chromosome 1, there are three additional such regions of clustered candidate genes: 82 – 86 Mb, 201 – 213 Mb and 247 – 252 Mb. Additional candidate regions were found on other chromosomes, including a 178 – 182 Mb region on chromosome 2, 154 – 158 Mb region on chromosome 3 and an 8 – 13 Mb spanning region on chromosome 7, to mention some of these (c.f. Figure 4.4). In addition to the *Vkorc1* region there are 20 additional such regions with clustered candidate genes in the genome of the rat (dense black bars highlighted with red squares in Figure 4.6B) that are of interest with regard to our experimental variables studied; notable warfarin induction and warfarin resistance genotype.

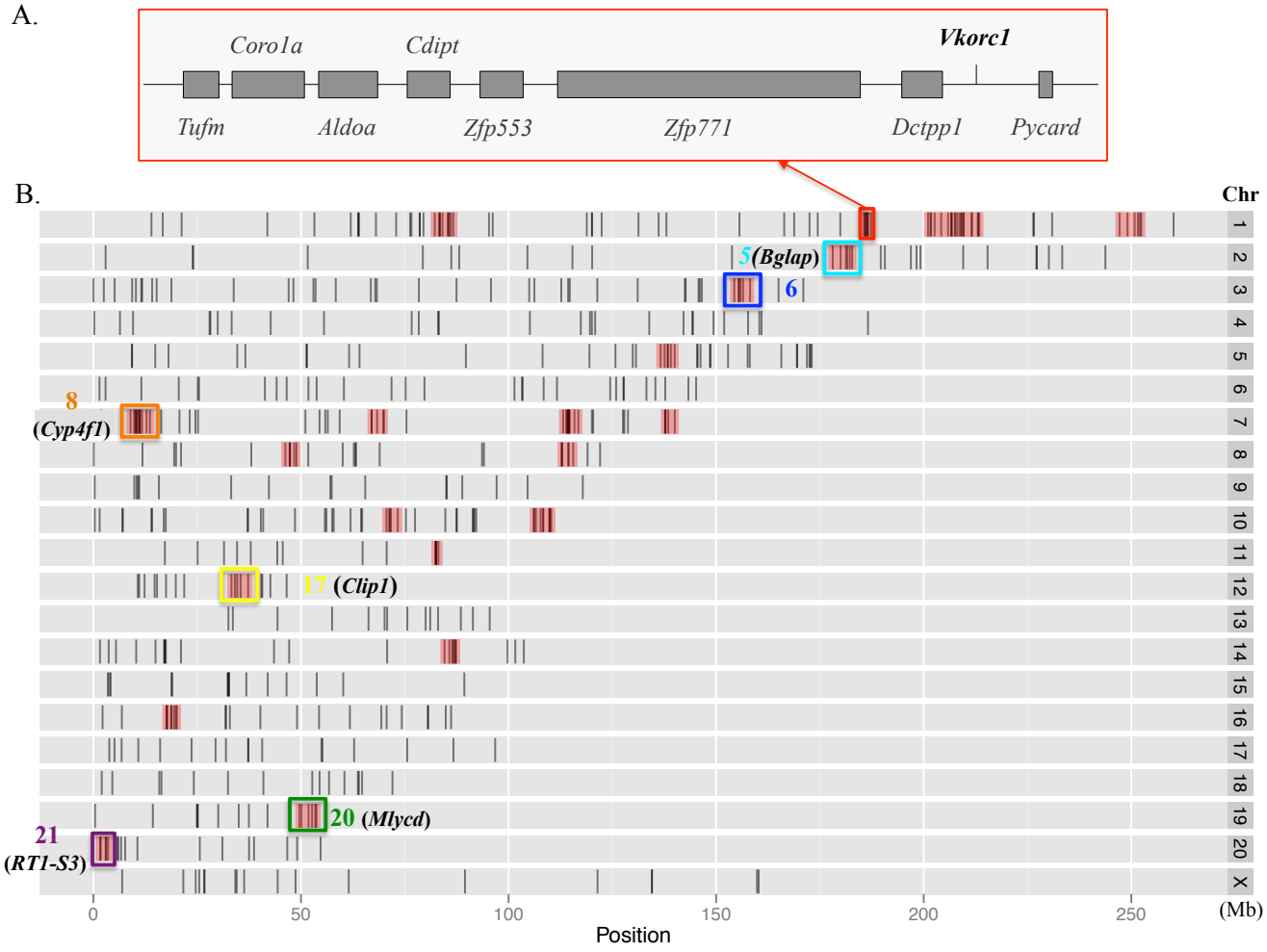


Figure 4.4 – 591 candidate target genes with warfarin related expression signals form 21 clusters along the genome. (A) the resistance gene *Vkorc1* locates in the 185-187 Mb region on chromosome 1 with 8 clustered candidate genes. (B) Regions of clustered genes are highlighted red, and regions with supports from SNP array analysis (Chapter 3) are marked with box of different colors (color code as in Figure 4.7), with region ID (as in Table 4.2) and corresponding candidate genes from SNP array analysis labelled beside.

4.3.1.4. Candidate genes have high connectivity in the co-expression network

We are interested in understanding the effect of network position has on the association of genes in our study context. Hub genes have higher connectivity that the average gene in the co-expression network, and thus, might have stronger effects on the trait (Ghazalpour et al. 2006; Benfey and Mitchell-Olds 2008; Barabasi, Gulbahce, and Loscalzo 2011). We calculated network connectivity (K_{all}, i.e. node degree), intramodular connectivity (K_{in}) and outmodular network connectivity (K_{out}) for each gene. In Table 4.4 we compare the average connectivity of trait-associated modules with the average connectivity of all modules in the co-expression networks. We observed that the former are generally larger than the latter. Moreover, we saw that the average connectivity values of trait-associated genes are higher than the background values. These observations indicate that hub genes in co-expression networks play important roles in mediating the adaption to warfarin as mediated by gene regulatory evolution.

Table 4.3 – Connectivity in co-expression network.

	Comp1			Comp2			Comp3			Comp4		
	K_{all}	K_{in}	K_{out}	K_{all}	K_{in}	K_{out}	K_{all}	K_{in}	K_{out}	K_{all}	K_{in}	K_{out}
All modules	95	20	75	121	12	108	80	16	64	86	13	73
Trait-associated modules	138	30	109	124	14	111	89	13	76	105	13	92
All genes	175	65	110	151	33	117	141	60	81	121	32	89
Trait-associated genes	255	87	168	181	44	137	184	83	101	143	39	104

Comp1: Resistant vs. Susceptible with warfarin induction; Comp2: Resistant vs. Susceptible without warfarin induction; Comp3: Warfarin Induction vs. Non-induction in Susceptible rats; Comp4: Warfarin Induction vs. Non-induction in Resistant rats. See Materials and Methods and Table1 for detailed experimental design in above four comparisons.

K_{all}: average total network connectivity; K_{in}: average intramodular network connectivity; K_{out}: average outmodular network connectivity.
All modules or genes: the average connectivity for all modules or genes.
Trait-associated modulesor genes: the average connectivity for modules or genes that correlate with traits (P-value < 0.05).

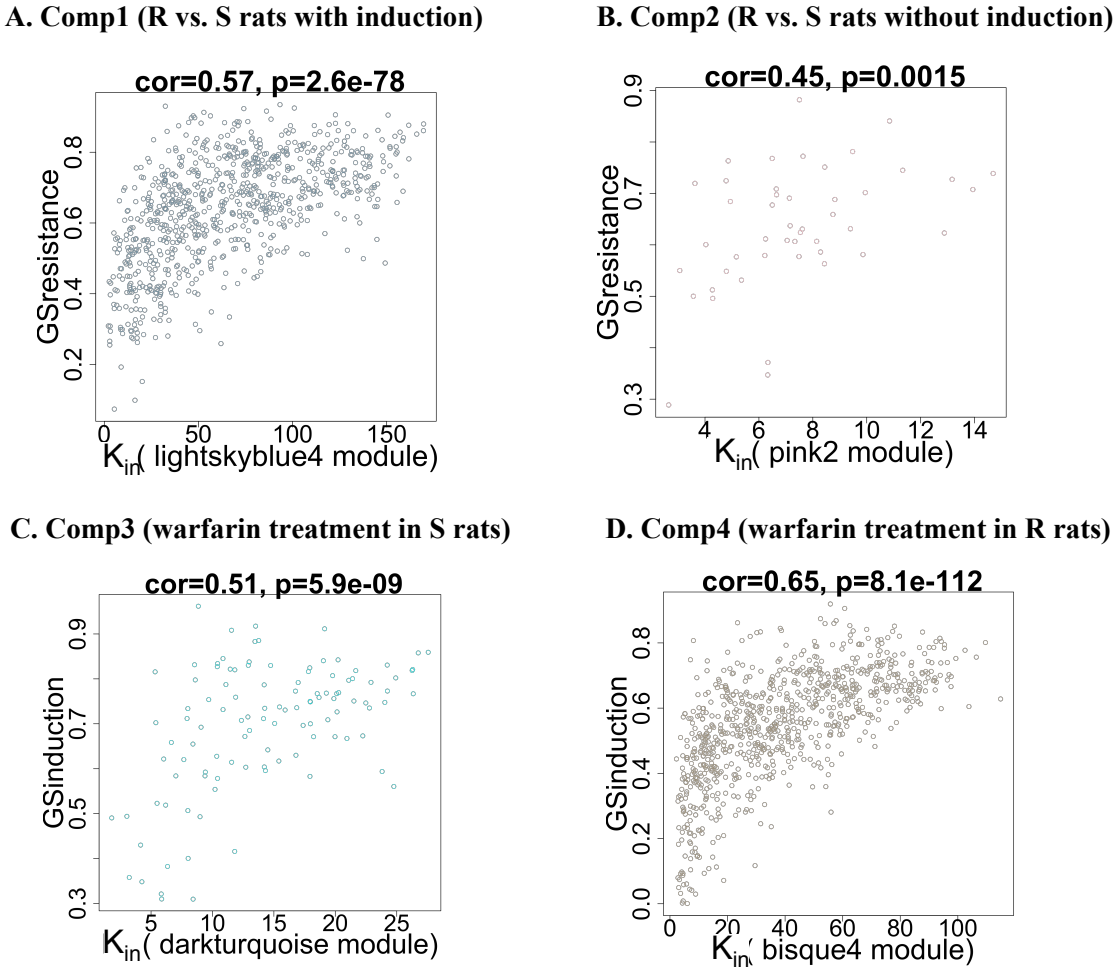


Figure 4.5 – Gene-trait relevance (GS) correlate with intramodular connectivity (K_{in}). The modules with high correlation with resistance or warfarin induction in each of the four co-expression networks are shown here. Other modules were not shown.

More specifically, we found that intramodular connectivity K_{in} significantly correlates with the relevance of genes to traits, such as can be seen from the lightpink module in Comp1, which contained the *Vkorc1* gene (Spearman correlation=0.57, P-value=1.7×10^{-60}). Within most trait-related modules genes with higher gene-trait relevance scores tend to have higher intramodular connectivity (Figure 4.5), which suggest that hub genes in co-expression network might have higher relevance to the resistance phenotype and/or warfarin treatment.

4.3.2. Gene-gene interaction information facilitates candidate identification

4.3.2.1. Gene ranking algorithm

The modified PageRank algorithm integrated gene-trait correlation measures into genetic network and gave high rank scores to genes if their interacting genes also had high rank scores (see Methods and Materials). It effectively prioritized genes highly relevant to traits, and would help decrease false positive signals. Thus we assigned gene ranks to the selected candidate genes to refine candidate list. Besides network topology, damping factor *d* is a user-defined parameter that would affect gene ranks, for it describes how far a gene feature would spread in the network (Fu, Lin, and Tsai 2006; Winter et al. 2012b). Google used 0.85 as the damping factor for web searches. In biological network, however, the damping factors should be lower since most genetic networks are much more sparse (Clune, Mouret, and Lipson 2013). We estimated dynamic damping factor ddf_i for each node in our network according to Fu et al's suggestion (Fu, Lin, and Tsai 2006) and used the average *ddf* as the damping factor for the whole network.

4.3.2.2. Properties of GGI network

First, we computed gene rank scores in gene-gene interaction (GGI) networks based on the information of protein-protein interaction and text-mining from STRING database (http://string-db.org/). We built four GGI networks according to the designed comparisons (Figure 4.1). The number of nodes and edges of GGI network for each comparison are summarized in Table 4.4. There is no much difference among GGI networks of different comparisons; the slight difference of gene numbers were due to the initial data cleaning process of microarray data.

Table 4.4 – Network statistics of GGI (gene interaction network) and CTD (comparative toxicogenomic database) network.

Statistics	Comp1		Comp2		Comp3		Comp4	
(N)	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
GGI	6573	109785	6617	111328	6568	109636	6616	111085
CTD	7474	203668	7516	205098	7477	203808	7513	204841

Nodes: genes. Edges: gene interactions.

4.3.2.3. CTD network

We also computed gene rank scores in CTD networks, which were built from a chemical-gene interaction database (Comparative Toxicogenomic Database, http://ctdbase.org/). Based on the rat chemical-gene interaction data we first retrieved warfarin related genes, and then their interactions with other genes if they were interacting with the same chemicals. The resulting warfarin related gene-gene interaction network data generated 4 CTD networks corresponding to 4 comparisons (Table 4.4).

In each network and comparison we obtained gene rank scores from both GGI and CTD networks, and then assigned them to our candidate genes. The GGI and CTD ranks for the 21 regions of clustered candidate genes after averaging rank scores within each region are shown in Table 4.2. Region 2 includes the known resistance gene *Vkorc1*, which has highest GGI rank; although *Vkorc1* itself didn't show trait related expression changes. Region 3 was ranked 1st when using the CTD data, and covers 10 Mb (201 – 213 Mb) on chromosome 1.

4.3.3. Candidate genes across different chromosomal regions share similar functions

To evaluate the functions of the 163 candidate genes from the clustered regions (Figure 4.6, Table 4.2), the GO (Gene-Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analyses were performed in DAVID (http://david.abcc.ncifcrf.gov). We obtained 149 function categories (GO/pathway/protein domain) and their enrichment score by comparing candidate genes from 21 clustered regions with the overall background gene list. Keeping the categories with enrichment P-values ≤ 0.05, 84 categories corresponding to 95 genes were grouped into 19 function clusters, each carrying genes sharing similar functions (Figure 4.6, Appendix 7).

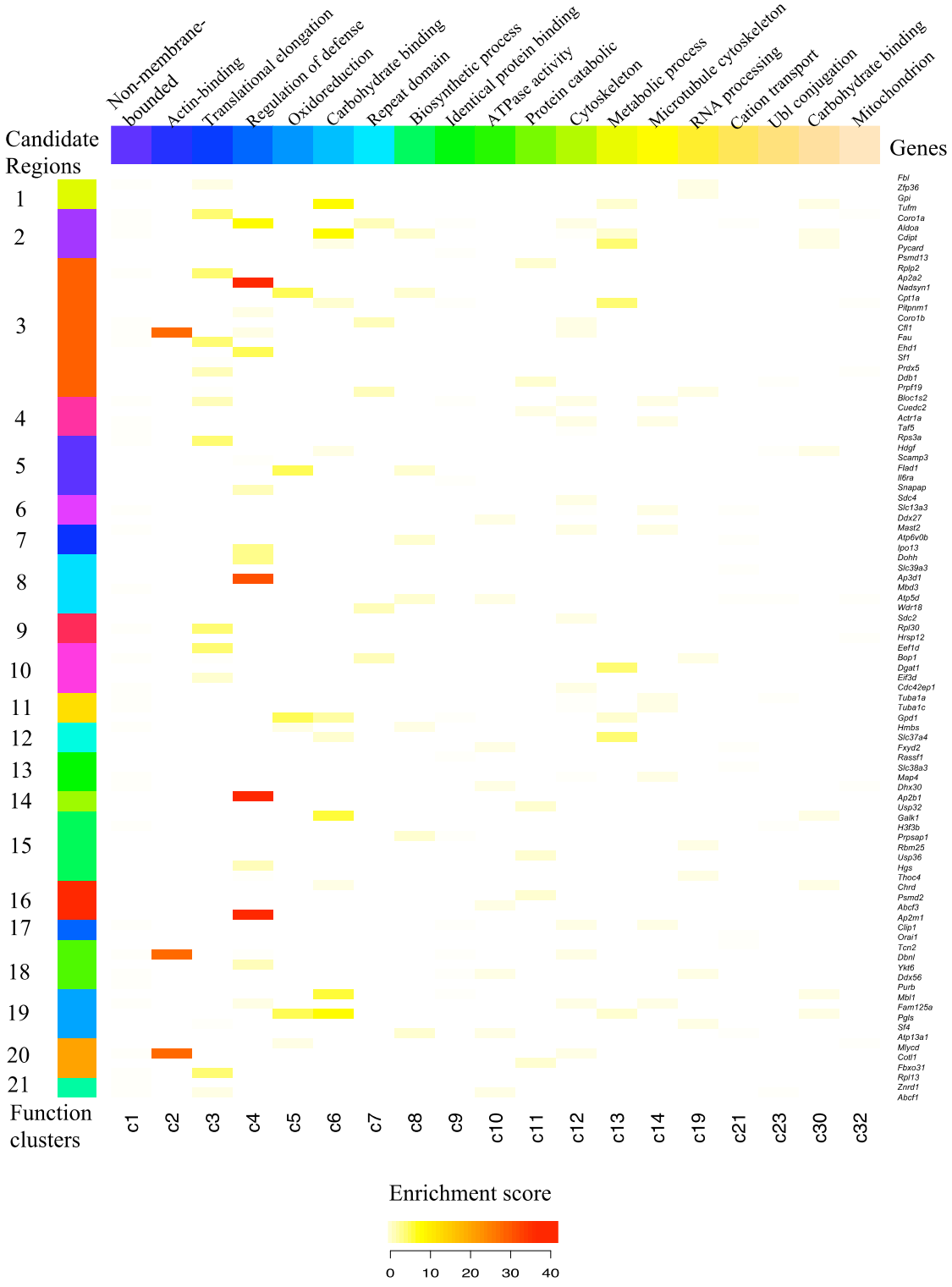

Non-membrane-bounded
Actin-binding
Translational elongation
Regulation of defense
Oxidoreduction
Carbohydrate binding
Repeat domain
Biosynthetic process
Identical protein binding
ATPase activity
Protein catabolic
Cytoskeleton
Metabolic process
Microtubule cytoskeleton
RNA processing
Cation transport
Ubl conjugation
Carbohydrate binding
Mitochondrion
Candidate Regions
Genes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Function clusters
c1
c2
c3
c4
c5
c6
c7
c8
c9
c10
c11
c12
c13
c14
c19
c21
c23
c30
c32
Fbl
Zfp36
Gpi
Tufm
Coro1a
Aldoa
Cdipt
Pycard
Psmd13
Rplp2
Ap2a2
Nadsyn1
Cpt1a
Pitpnm1
Coro1b
Cfl1
Fau
Ehd1
Sf1
Prdx5
Ddb1
Prpf19
Bloc1s2
Cuedc2
Actr1a
Taf5
Rps3a
Hdgf
Scamp3
Flad1
Il6ra
Snapap
Sdc4
Slc13a3
Ddx27
Mast2
Atp6v0b
Ipo13
Dohh
Slc39a3
Ap3d1
Mbd3
Atp5d
Wdr18
Sdc2
Rpl30
Hrsp12
Eef1d
Bop1
Dgat1
Eif3d
Cdc42ep1
Tuba1a
Tuba1c
Gpd1
Hmbs
Slc37a4
Fxyd2
Rassf1
Slc38a3
Map4
Dhx30
Ap2b1
Usp32
Galk1
H3f3b
Prpsap1
Rbm25
Usp36
Hgs
Thoc4
Chrd
Psmd2
Abcf3
Ap2m1
Cllp1
Orai1
Tcn2
Dbnl
Ykt6
Ddx56
Purb
Mbl1
Fam125a
Pgls
Sf4
Atp13a1
Mlycd
Cotl1
Fbxo31
Rpl13
Znrd1
Abcf1
Enrichment score
0
10
20
30
40

Figure 4.6 – Heatmap of candidate genes' enrichment across functional clusters. Candidate target genes form 21 clusters along the genome. Function clusters are identified based on gene ontology and pathway analysis. Genes sharing similar functions are distributed at different regions along the genome.

4.3.4. Candidate genes in clustered regions are connected in gene-gene interaction networks

Among the 591 genes with detected warfarin-related expression signals, 163 genes are clustered into 21 chromosomal regions along the genome (Figure 4.6B). Based on Gene Ontology annotation analysis, we noticed some genes from different chromosomal regions share similar functions (Figure 4.6). This observation inspired the questions of what are the relationships among these 163 candidate genes? It is expected that one or two genes in each region would be directly affected by warfarin, and other genes in the same region with expression changes are physically linked to these 'causal' genes. Here we depicted the gene interactions among these 163 candidate genes in 21 genomic regions based on the gene-gene interaction data downloaded from the STRING database. Nodes in Figure 4.7A are genes. Two genes are connected using red line if they are interacting with each other; and the names are shown for these genes. If there is no interaction between two genes, but they are located in the same chromosomal region, they are connected using gray lines. And the genes with only chromosomal proximity connections with others were indicated with small gray nodes without name shown.

Corresponding to the 21 chromosomal regions, we observed 21 clusters of 163 genes in Figure 4.7A (the number besides each cluster represents the region ID as in

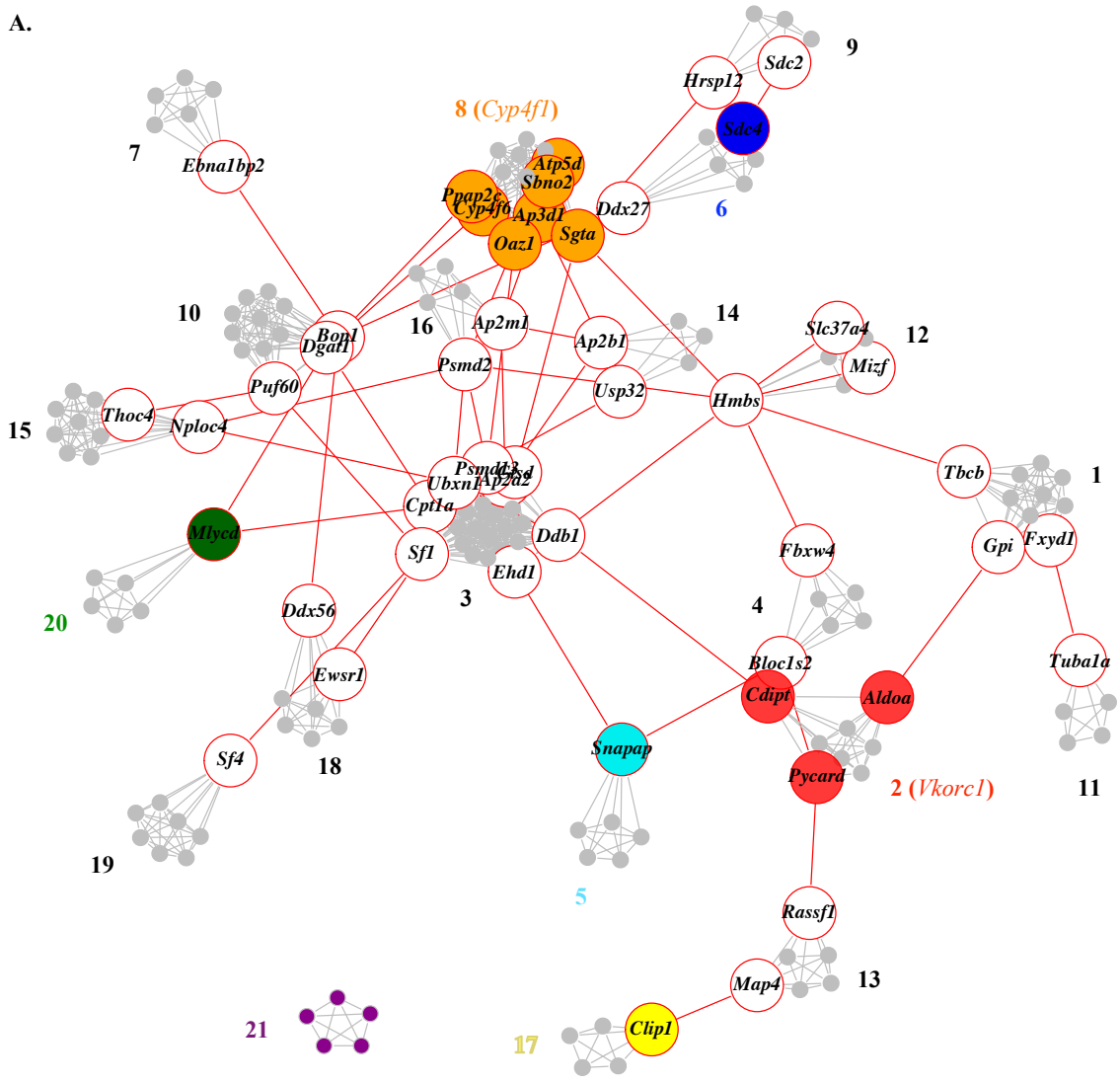
Table 4.2). We saw that most (20) clusters are connected except one is isolated. Consistent with our expectation, most connections are made through one or two genes in each region/cluster. The region *Vkorc1* located in (region 2, red) has three genes interacting with other genes. As mentioned above, *Vkorc1* itself has no expression signals; but the expression profiles of other genes in this region were probably affected by its sweep effect.

Another two regions (region 3 and region 8) gained our attention that they have multiple genes (8 and 7) connected to other clusters in terms of gene interaction. Region 3 is located at chromosome 1 (201-213 Mb), which is right next to the sweep region of *Vkorc1*. But it is too early to say the expression changes of these genes are warfarin relevant. Region 8 is at chromosome 7 (8-13 Mb, orange), which is also detected in previous SNP array analysis (Chapter 2).

From the NetGWAS analysis of rat SNP array data, we have identified 87 candidate genes (Appendix 4, c.f. Chapter 3). We compared these 87 candidate genes detected based on genetic variations with the 163 genes with expression signals related to warfarin (Appendix 6). 7 chromosomal regions with supports from both SNP array and expression analyses were highlighted along the chromosomes (Figure 4.6B), as well as in gene interaction networks using the same color code (Figure 4.7A).

26 candidate genes from SNP array and 49 target genes with expression signals are located in these 7 regions. So we pooled these genes and drew gene interaction network specifically for them (Figure 4.7B). As in Figure 4.7A, nodes are genes; region ID is labeled beside each cluster with the same color code. Two genes are connected

using red line if they are interacting with each other (big nodes with gene name shown); two genes are connected using gray line if they are in the same chromosomal region. Genes only share chromosomal proximity with others are small gray nodes.



B.

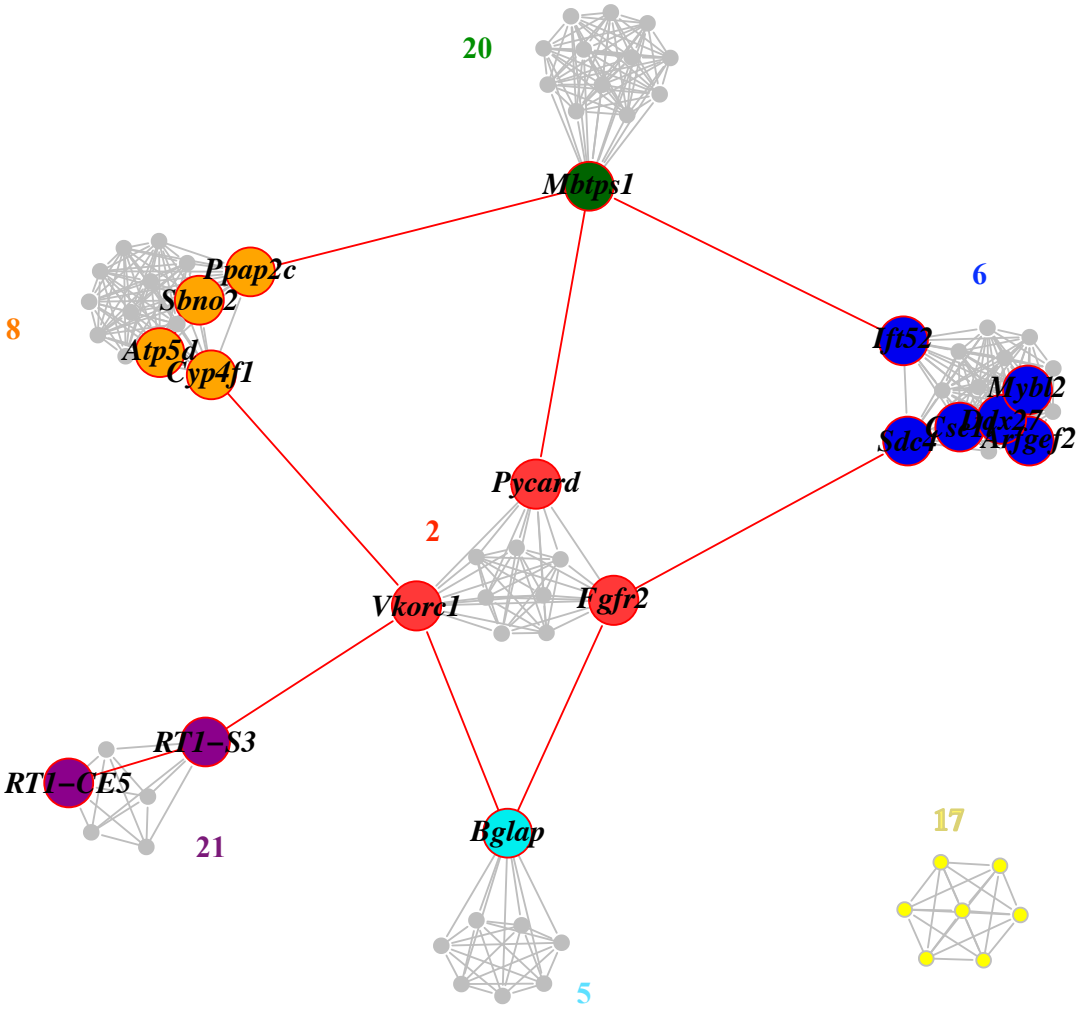


Figure 4.7 – (A) Gene-gene interactions among 163 target genes clustered in 21 regions along the genome. (B) Gene-gene interactions among 73 genes in 7 regions supported by both SNP array and Microarray analysis. Node: gene. Red line: gene-gene interaction. Gray line: genes in the same chromosomal region. Gene names are not shown for genes that don't interact with others. Region ID is labeled besides.

6 out of 7 regions are connected and as expected *Vkorc1* region is the center. *Vkorc1* is interacting with *Cyp4f1* gene in region 8, *Bglap* gene in region 5 and *RT1-S3* gene in region 21. *Cyp4f1* gene belongs to cytochrome P450 gene family, in which most members were found to be drug-metabolizing enzymes (Thomas 2007). Moreover, *Cyp4f1* is the ortholog of *CYP4F2* gene in human, which recently has been identified to be associated with warfarin dosage variance (Pautas et al. 2009; Takeuchi et al. 2009).

Bglap gene encodes a vitamin K dependent protein that is involved in vessel calcification and bone mineralization (Suttie 1993; Danziger 2008). As we will mention in Chapter 6, we detected that *Bglap* gene is associated with both arterial calcification and warfarin resistance. Thus it might be a good starting point to investigate the fitness costs of warfarin resistance in terms of arterial calcification (Kohn, Price, and Pelz 2008).

The region 6 at chromosome 3 (154-158 Mb) was observed with relatively high association with warfarin resistance in SNP array analysis (Chapter 3), and was also detected with expression signals. But which gene in this region is related to warfarin was a puzzle to us. Here we saw that region 6 is connected with the *Fgfr2* gene, which is a fibroblast growth factor receptor and is 2 Mb away from *Vkorc1* on Chromosome1. Within this distance, *Fgfr2* should be under the sweep effect of *Vkorc1*; the potential hitchhiked variant in *Fgfr2* and its interaction with region 6 might be one explanation for the detected association signals at region 6. But it is also possible that *Fgfr2* and genes in region 6 are functional relevant to warfarin resistance, since *Fgfr2* plays important roles in embryonic development and tissue repair, especially for bone and blood vessels. Moreover, in human a fibroblast growth factor binding protein gene (*FGFBP2*) exhibited

relatively high association with warfarin dose (Cooper et al. 2008). Considering the fitness cost of reduced growth rate in resistant rats and the 'blood-thinning' effect of warfarin, this *Fgfr2* gene might provide another path for us to explore the connections with warfarin.

The region 6 and region 8 are both connected to region 20 by gene *Mbtps1* (membrane-bound transcription factor peptidase, site 1). It might be another interesting gene for further examination.

4.3.5. Summary of top candidate genes

All the candidate genes were shown in Appendix 6. Table 4.2 listed the 21 candidate regions with clustered targets (163 genes) along the genome. Figure 4.7A depicted the gene-gene interactions among these 163 genes. 95 genes of them were grouped into 19 function clusters based on function annotation analysis, with their enrichment scores across function clusters shown in Figure 4.6. Here the functional sharing pattern is different from the pattern of similar functions sharing among candidate genes identified from the genomic variant data (c.f. Chapter 3). Genes located within the same candidate regions don't necessarily share similar functions; they are detected as their expression changes associated with warfarin related traits.

The five genes at region 2 (185 – 187 Mb on chromosome 1), where the resistance gene *Vkorc1* locates, exhibited different function profiles. However, enrichment of similar function clusters was observed across different chromosomal regions. For example, 12 out of 21 candidate regions carry genes with function categories in cluster 4,

which are related to regulation of defense response and membrane coat (Appendix 7). Function categories in cluster 3 shared by genes in 11 candidate regions, are annotated as translation elongation and protein biosynthesis.

In human, in addition to *VKORC1*, there are two genes identified as predictors for warfarin dosage: *CYP2C9* and *CYP4F2*; both of them belong to cytochrome P450 gene family (Rost et al. 2004; Takeuchi et al. 2009). Most members in this family have been recognized as drug-metabolizing enzymes (Thomas 2007). More specifically, some P450s in the Cyp2c, Cyp2e and Cyp3a family have been detected with differentiated expression related to anticoagulant resistance (Imaoka, Hashizume, and Funae 2005; Markussen et al. 2008b; Markussen et al. 2008a). We noticed there are two P450 genes *Cyp2t1* and *Cyp4f6* that map in regions with label 1 and 3 on chromosome 1 and region 8 on chromosome 7, respectively. These share functional attributes including the oxidation-reduction cycling with *Vkorc1* and ion binding with other P450 genes.

As shown in Figure 4.6 and Table 4.2, the P450 gene *Cyp4f1*, detected as a candidate gene from NetGWAS (Chapter 2), also maps on chromosome 7 region 8, close to the *Cyp4f6* gene. *CYP4F2* gene is a vitamin K_1 oxidase (McDonald et al. 2009). A comparison study between human and rat showed that both *CYP4F2* in human and its ortholog *Cyp4f1* in rat produced 19/20-HETE (hydroxyeicosatetraenoic acids) (Imaoka, Hashizume, and Funae 2005). As the region on chromosome 1where *Vkorc1* is located, the region of 8 – 13 Mb on chromosome 7 containing *Cyp4f6* and *Cyp4f1* genes are interesting for further examination based on signals from both expression and SNP array studies.

In addition, we observed strong SNP-resistance association signals at the region of 154 – 158 Mb on chromosome 3. And here the microarray analysis revealed expression signals at this region again (region 6), which can hardly be observed by chance. As shown in Figure 4.7B, genes in this region are interacting with the Fgfr2 gene, which is under the sweep effect of *Vkorc1* or it might be functional relevant. The next challenge for us would be locating the potential causal gene in this area.

Overall, other than the known resistance gene *Vkorc1*, we listed 20 regions with clustered candidate genes from microarray analysis; 7 of them are paralleled with supporting evidences from NetGWAS of SNP array (Figure 4.6, Figure 4.7). The candidate genes in Table 4.2 suggest a starting list for further exploration.

4.3.6. Traditional analysis identifies differentially expressed genes

Before conducting gene co-expression network analysis, we performed traditional gene expression analysis to detect genes differentially expressed between two phenotypes (resistant vs. susceptible) and warfarin treatment (induction vs. non-induction). Applying cutoffs of delta values suggested by permutation based estimation of false discovery rates (FDR) (Figure 4.2), we obtained 55 differentially expressed genes considering either resistance phenotype or warfarin induction (data not shown). 9 of 55 differentially expressed genes were on our candidate target list identified from co-expression analysis; 16 of them were located in the regions of clustered candidate targets. But without the co-expression analysis results, this list of 55 genes itself gave us limited information about the potential candidate genes.

We also obtained another list of differentially expressed genes (101 genes, data not shown) considering both resistance phenotype and warfarin induction using P-value ≤ 0.05 as a threshold. 26 of 101 differentially expressed genes matched with our candidate target genes; 19 of them were located in the clustered gene regions. Though these differentially expressed genes identified using traditional approach didn't provide clear hint about candidate genes for next validation, they are kept for comparison purpose with the co-expression network analysis results.

4.3.7. *cis*-eQTLs (quantitative trait loci) are enriched in candidate genes

Gene expression variation is abundant in species from yeast to human, and plays important role in connecting genotypic variation to phenotypic variation (Morley et al. 2004; Gilad, Rifkin, and Pritchard 2008). How would the nucleotide change affect phenotype by modifying gene expression profiles? By expression quantitative trait loci (eQTL) mapping, we could identify genetic factors involved in gene regulation. As eQTL studies have demonstrated its promise in human biomedical and adaptation research since 2004 (Morley et al. 2004; Schadt et al. 2005; Gilad, Rifkin, and Pritchard 2008; Veyrieras et al. 2008; Kudaravalli et al. 2009), we try to detect expression associated genetic variations by examining both the SNP array (c.f. Chapter 3) and microarray data of the wild-derived NW rats. It has been suggested that eQTLs are targets of recent positive selection in human (Kudaravalli et al. 2009). Here in rat populations under strong selection of warfarin, we expect to observe enrichment for eQTLs in our candidate genes rather than the genetic background.

We tried to identify *cis*-eQTLs that the SNPs are located within 100 kb of the gene, and also are associated with gene expression profiles based on standard regression test (Kudaravalli et al. 2009). With a threshold of P-value ≤ 0.05, we gained 506 SNPs as *cis*-eQTLs, which correspond to 550 genes. In addition, we selected a smaller set of SNPs (19 eQTLs) with significance level threshold ≤ 10^{-4} as suggested (Kudaravalli et al. 2009).

With the identified *cis*-eQTL, we tested whether they were enriched in our identified candidate genes. Here we considered several candidate gene list, including the 591 candidate genes based on expression network analysis (Appendix 6), 101 or 55 differentially expressed genes using traditional expression analysis (data not shown), and 87 candidate genes based on SNP array analysis (c.f. Chapter 3). As expected, eQTLs are indeed enriched (P-values ≤ 0.01) in both the 87 candidate genes from SNP array and 591 candidate genes from expression analysis (Table 4.5). But they are not overrepresented in the 55 differentially expressed genes identified using traditional approach. These enrichment results are consistent with our hypothesis that warfarin selection lead to more expression variations, thus support our identified candidate genes.

Table 4.5 – *cis*-eQTL enrichment test in candidate genes

	Expression analysis (Microarray)			NetGWAS(SNParray)
Tested gene list	591 targets[α]	101 differentially expressed genes[β]	55 differentially expressed genes[β]	87 candidate genes (c.f. Chapter 3)
eQTL No.	36	7	4	8
P-value (=hits of eQTL)	0.001	0.049	0.099	0.011
P-value (>hits of eQTL)	0.002	0.036	0.058	0.005

[α] candidate target gene list based on expression network analysis.
[β] differentially expressed gene list based on traditional expression analysis.

4.4. Discussion

Natural selection is the key of Darwin's explanation of biodiversity and adaptation. What is actually being selected at the molecular level? One is amino acid change of protein sequences, and another is the expression levels' or patterns' alteration. Though positive selection on coding region has been intensely studied, the adaptive expression changes have been paid less attention. Using rat resistance toward anticoagulant drug warfarin, we demonstrated that there is a huge response of expression variations (~ 600 candidate genes) related to warfarin. The candidate genes form 21 clustered regions along the genome, ranging from 5-22 genes with average of 8 genes in each cluster. These candidate clusters implies the existence of 'true' warfarin related genes and the potential hitchhiking effects on their flanking regions. In addition, our study suggested that the strategy of network-guided approach should be useful for complex systems, like complex diseases in human.

4.4.1. Building gene co-expression network is an effective way to detect trait relevant expression variations

Although our microarray dataset is incomplete and error-prone, it provides us an access to genomic changes embedded in expression variations response to warfarin stimuli. To understand the source of expression variations, networks with thousands of interacting genes that orchestrate expression need to be considered (Wittkopp 2007). The success of using genetic network to predict genes involved in specific pathways for either

unicellular organisms or animals has been demonstrated before (Lee et al. 2008). Here we constructed genome-wide co-expression networks, but aimed to identify genes specifically related to warfarin selection.

Modularity is an innate property of many genetic networks and a major contributor to organisms' evolvability (Clune, Mouret, and Lipson 2013). Highly connected co-expressed modules could be consistent with co-regulated genes related to certain function or adaptive to certain environmental change (Ghazalpour et al. 2006; Wittkopp 2007). Similar to the engineering concept, modularity maintains integrity, security and reliability. It might be evolved because of common challenges in rapidly changing environments or as a byproduct of selection reducing costs of connections in genetic network (Clune, Mouret, and Lipson 2013). Here we detected modules of genes with similar expression profiles that are associated with warfarin treatment or resistance phenotype.

Network connectivity predicts gene essentiality. Gene essentiality is a general but important parameter, which would not only aid in targeting candidate genes in warfarin selection. It was reported that gene connectivity of probabilistic network could predict gene essentiality in yeast, worm and mouse (Lee et al. 2008). In our analysis, the positive correlation between intramodular connectivity of genes and their extents to trait association supported this idea in co-expression network (Figure 4.5).

The rats' expression data and gene connectivity measures in co-expression network are accessible to the community for interests in other pathways or diseases.

The interactions between genes would help identify signals which otherwise might be missed using traditional single-locus analysis, especially for those loci with moderate or small single-locus effect on phenotype (Liu et al. 2010).

4.4.2. *CYP450* genes are not overrepresented in candidate genes involved in warfarin resistance

Previously, Markussen et al reported that cytochrome P450 genes play important role in 4-hydroxycoumarin-based anticoagulant resistance by examining their hepatic expression profiles upon bromadiolone administration (Markussen et al. 2003). Cytochrome 9450 gene family usually includes 50-80 genes and human genome contains ~ 60 genes. Genes in this superfamily that encode enzymes to metabolize known endogenous substrates are phylogenetically stable; and other unstable genes function in detoxification of xenobiotic compouds (Thomas 2007). Here we find that, in statistical terms, the *CYP450* genes are not overrepresented in our candidate gene list (hypergeometric test, P-value of more than 2 *CYP450* genes as candidate genes is 0.900). And the study on the long-term evolutionary stability of *CYP450* genes found that sites of selection in these xenobiotic related genes are mostly associated with changes in protein structure rather than modification in expression regulatory regions (Thomas 2007). Thus, the importance of *CYP450* genes in expression variation induced by anticoagulant drugs as previously inferred would be more like a result of the experimental setup, where only these have been assayed and thus, necessarily some will emerge. But we still cannot rule out the possibility that *CYP450* genes did play key role in expression variation associated with bromadiolone, which is a second-generation rodenticide, belonging to 4-

hydroxycourmarin compounds. Further examination based on expression data with bromadiolone treatment is needed to understand the genetic architecture (c.f. Future direction in Chapter 6).

Chapter 5

Calu and other candidate genes are associated with warfarin resistance in wild Norway rats as revealed by population structure analysis and NetGWAS

Abstract

Drift plays important role in human population because of strong demographic effects, thus adaptive signals might be obscure across different populations. However, we expect to find candidate genes with discernible adaptive signals in rat populations since they have experienced strong selection of warfarin. We performed the network-guided GWAS of the rat 10k SNP array data (SNP array II) in multiple wild rat populations. 46 samples were chosen from genetically homogeneous virtual populations inferred by population structure analysis of ~ 600 rats. Following the NetGWA analysis, genes with multiple supporting evidences across populations were reported as top candidates. We have collected the genotype data for one candidate gene *Calu* (Calumenin) in ~ 600 rats, and observed significant associations with warfarin resistance in 7 natural populations, compared to the observation of significant associations in 13 natural populations for *Vkorc1* gene. The significant associations for *Calu* and *Vkorc1* gene remain in inferred virtual populations. Given *Calu*'s regulation role in the vitamin K-dependent carboxylation pathway, the observed association in multiple natural populations suggests *Calu* is another warfarin-related gene merited further investigation. The result also provides foundation for examining the effect of gene interactions on adaptive consequences.

5.1. Introduction

We have obtained a list of candidate genes based on GWAS (c.f. Chapter 3) and gene expression analysis (c.f. Chapter 4). However, both the SNP array and microarray

data were collected from samples of a wild-derived population NW. To evaluate the candidate list and to search for additional candidate genes in natural populations, we collect more genotype data of wild rat samples based on the same SNP array, and for distinguishing purpose we called it SNP array II.

In these wild populations, how to choose samples for SNP array experiment is a problem. We have sampled ~ 700 rats from 19 farms in a resistant area at north-western Germany (Figure 5.1). Usually the population is naturally determined by the geographical origin of samples. But rats from different farms are not necessarily genetically differentiated; on the other hand individuals from the same location could be genetically structured due to unidentified barriers to gene flow (Evanno, Regnaut, and Goudet 2005). With this recognition in mind, we analyzed the population structure for the ~ 700 rats before we selecting samples for SNP array experiment and genomic association analysis.

We collected genotype data of microsatellites and presumably neutral SNPs for ~700 rats. Population structure analysis is implemented in two commonly used softwares: STRUCTURE and INSTRUCT. It is always a challenge to decide the optimal number of clusters (K) when identifying genetically homogeneous groups of individuals. We used different criterions to the best of our knowledge to detect K, which would be a useful reference for future population structure studies.

The following GWAS and network-guided GWAS were conducted the same as described in Chapter 3. Combining genetic variation data and expression data in multiple populations, we refined the candidate gene list and summarized them for further validation. We have already collected genotype data and conducted association tests for

one candidate gene *Calu* in multiple natural populations, which shows the relationship of *Calu* and warfarin resistance, and is also the foundation for future gene interaction analysis.

5.2. Materials and Methods

5.2.1. Rat (*R. norvegicus*) samples

742 wild *Rattus norvegicus* samples were sampled from 36 farms located in north-western Germany with reported anticoagulant resistance (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). Farms with small sample size (< 10) were disregarded from further analysis, resulting in 691 rats from 19 farms (Appendix 1). We used the abbreviation names for 5 populations (WU-pop24, KB-pop11, TH-pop23&16, SP-pop20, LH-pop4f (a control population from non-resistant area)) as mentioned in previous study (Kohn, Pelz, and Wayne 2000). Other populations were called by their farm numbers as used in previous study (Kohn, Pelz, and Wayne 2003). 618 samples of above 691 rats were assigned with warfarin resistant/susceptible phenotypes, which were physiologically determined with a blood clotting response (BCR) test (Thijssen 1995; Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). For other three second-generation rodenticides, we also obtained the bromadiolone and the coumatetralyl resistant/susceptible phenotype depending on the BCR test (Thijssen 1995; Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). The resistant level R% for each population is calculated using the number of resistant samples divided by the number of samples with warfarin related phenotype tested.

5.2.2. Microsatellites and SNP data for structure analysis

We randomly selected 8 SNPs on the intergenic region of the rat genome. 4 of them were located in the gene desert region (the nearest gene were 1 Mb away).

First, we conducted PCR and sequencing to obtain the genotypes of 10 samples for each SNP. The primers covering the identified SNP sites (product length: 1000-1200 bp) (Appendix 2) were designed in Primer3 (http://frodo.wi.mit.edu/primer3, access May 2012). For each DNA sample, a final reaction volume of 13.5 µl buffer with 1.5 µl genomic DNA was prepared for PCR. In general, PCR was preformed with an initial denaturation of 94°C (2.5 minutes), followed by 33 cycles of denaturation at 94°C (0.5 minutes), annealing at 57.5°C (40 seconds), and extension at 72°C (1 minute). The final extension was conducted at 72°C (5 minutes). For the primer dimmer cleanup, we added 3.5 µl ExoSAP-IT to 7 µl PCR product, and incubated at 37°C for 15 minutes and then inactivated ExoSAP-IT at 80°C for 15 minutes. PCR product was sent for both forward and reverse sequencing using ABI 3730 xl sequencer (DEWALCH, Houston, TX). The sequenced contig were assembled and edited using Phred/Phrap/Consed (http://www.phrap.org/) (Ewing et al. 1998; Gordon, Abajian, and Green 1998).

With 10 reference genotypes for each SNP, we genotyped the 691 rat samples from 19 natural populations by High Resolution Melting (HRM) test (Wittwer et al. 2003). The HRM PCR Kit was purchased from QIAGEN (http://www.qiagen.com, access Jun 2012). The primers covering the identified SNP sites (product length: 100-130 bp) for HRM genotyping (Appendix 2) were designed in Primer3 (http://frodo.wi.mit.edu/primer3, access May 2012). For each rat DNA sample, 0.65 ul

genomic DNA was added to reaction volume of 10 ul, which includes 5 ul HRM PCR Master Mix, 2.95 ul RNase-free water and 1.4 ul primer Mix. The HRM analyses were conducted on the Rotor-Gene Q, with an initial PCR activation of 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 10 seconds, annealing and extension at 55°C for 30 seconds; the following HRM were performed in a temperature range from 75°C-90°C with 0.1°C increments at each step. The temperature range for SNP S696656 is from 67°C-82°C as its Tm is around 75°C. The genotypes for each rat sample were called using two combined clustering methods with 10 reference genotypes in R (Team 2012). As recommended by (Reja et al. 2010), we carried out the linear discriminant analysis (LDA) to cluster unknown samples into groups after learning the features of known genotypes using the R package MASS (Venables and Ripley 2002). The two approaches were different in the curve scaling and normalization methods (simple scaling of the fluorescence rate from 0 to 100 and also the Levenberg-Marquardt method) (D.W.Marquardt 1963).

We combined the genotype data of 8 SNPs and 5 microsatellites from previous study (Kohn, Pelz, and Wayne 2003) for the population structure analysis. 5 neutral microsatellite loci: *D2Rat31*, *D10Rat6*, *D13Rat18*, *D14Rat15*, and *D17Rat38* were distributed on rat chromosomes 2, 10, 13, 14 and 17 respectively. Samples from LH population were excluded since they were sampled from a non-resistant area about 300 km away from the main resistant area, resulting genotype data for 603 rat samples.

5.2.3. Structure analysis

Based on the SNP and microsatellites data of 603 rats sampled from 18 farms, we analyzed the population structure using two Bayesian methods: Structure 2.3.3 (Pritchard, Stephens, and Donnelly 2000) and INSTRUCT (Gao, Williamson, and Bustamante 2007).

Parameter setting*:* In Structure the parameter settings were as follows:

1. Admixture model was applied since our rat samples may have mixed ancestry among different populations (Non-admixture model was tested and gave similar results).

2. Correlated allele frequency model was chosen with the assumption that frequencies of neutral loci are likely to be similar among different populations (independent allele frequency was tested and gave less convergent Ln Pr(X|K)).

3. Without given pre-defined population information, population structure was inferred from genetic information. (Estimation with pre-defined population information (given by the farm location that each sample was sampling from) resulted in large fluctuate values of α (1.7-0.5), which indicates poor separating populations).

4. Lambda, the parameter of the distribution of allelic frequencies, was set to 1 as suggested.

5. Burn-in iterations of 10,000 followed by 10,000 MCMC (Markov Chain Monte Carlo) repetitions were set as recommended (Pritchard, Stephens, and Donnelly 2000).

In INSTRUCT, we ran 5 independent chains of 200000 iterations with 100000 burn-in iterations.

Choosing optimal K*:* To appropriately specify the optimal number of the estimated cluster (K), we ran K = 1 to 18 (the number of sampled locations) and repeat 20 runs for each K in STRUCTURE AND 5 chains in INSTRUCT. We examined several criterions to choose the optimal K. 1) the posterior probability Ln Pr(X|K) (Pritchard, Stephens, and Donnelly 2000). Optimal K was selected when posterior probabilities Ln Pr(X|K) converged for larger K and the dirichlet parameter α for the degree of admixture settled down to be plateaus. 2) Evanno et al. developed another statistic, the second order rate of change of Ln Pr(X|K) weighted by the variance - delta K. The peak of delta K was used for detecting the true K (Evanno, Regnaut, and Goudet 2005). The analysis was implemented using Structure Harvester (Earl and vonHoldt). 3) We also estimated the delta F_{ST} (similar to delta K) and the average maximum correlation coefficient for each K using CorrSieve in R to assess the optimal K (Cockram et al. 2008; Campana et al. 2011). 4) We calculated the average pairwise Euclidian distance between matrices of predicted allelic frequencies using R scripts with the algorithm described by (Camus-Kulandaivelu et al. 2006). Then with the distance matrix we built the Neighbor-Joining tree using the R package ape. The NJ tree would also help decide the suitable K.

With the chosen K, we performed the permutation in CLUMPP (Jakobsson and Rosenberg 2007) to minimizing the inconsistency across replicates resulting from ‘label switching’ or ‘genuine multimodality’. The resulted individual membership Q matrix were visualized using DISTRUCT (Rosenberg 2004).

We calculated a similarity index to combine the virtual population grouping results from STRUCTURE and INSTRUCT as described by (Saïdou et al. 2009). 534 samples with agreement of population membership between STRUCTURE and INSTRUCT were kept for further analysis. And here we called the assigned population membership based on structure analysis as virtual populations.

5.2.4. Select samples for rat 10k SNP array

We selected 46 samples based on their natural and virtual population (VP) structure, considering both K=3 and K=6. VP3.X represents the virtual population group number X according to K=3; similarly VP6.X represents the virtual population group number X according to K=6. For SNP array experiment, 18 samples from pop20, 20 samples form WU, 5 samples from TH and 3 samples from other 3 populations were mainly belong to VP3.1 and VP3.3, or VP6.1, VP6.3 and VP6.6.

The population structure of all populations and the above 46 samples for SNP array experiment were shown in Figure 5.1.

5.2.5. Analysis of SNP array data

Genomic DNA of 46 rats were isolated using the classical phenol:chloroform extraction method from liver tissues. Genome scale SNP data of the rat samples were collected using the rat 10K array purchased from Affymetrix (c.f. Chapter 3. Methods and Materials). Quality control tests were performed by Baylor Genomic & RNA Profiling Core and Vanderbilt University for us. The microarray center at Vanderbilt University conducted the genotyping experiment on our behalf. The genotypes were

called following data normalization and quality control using GTGS (Affymetrix GeneChip Targeted Analysis Software). With 10847 SNP sites from Rat 10k array, filtering out 493 failed SNP sites and 3594 non-informative SNPs, we obtained 6760 informative SNPs for 46 wild rats. We performed the association tests and population genomic tests on all the 6760 informative SNPs.

We followed the same procedure as describe in Chapter 3 to analyze the SNP array data. First, conducted the traditional GWAS using P-values and Bayes Factors as genotype-phenotype association measures. Then mapped SNPs to genes, and computed gene ranking in gene interaction network based on STRING database (http://string-db.org/, accessed July 2012). The modified PageRank algorithm was used for calculating gene ranks. Considering the effect of network topology, we corrected the gene ranks by random rank scores (c.f. Figure 3.4). The top ranked genes were selected for function annotation analysis. This is a validation of previous identified candidate genes and also a new discovery process.

5.2.6. SNP discovery, genotyping and association tests for *Calu* gene

8 primer pairs of *Calu* gene covering the 7 exon regions as well as the flanking intronic regions were designed to amplify about 1kb sequence of each region. To evaluate the potential hitchhiking effect of warfarin resistance, we also detected the SNPs on the intergenic flanking region of *Calu* gene. All the primer pairs were designed using Primer3 (http://frodo.wi.mit.edu/primer3, access May 2011) and are available in Appendix 2.

32 rat samples in the wild-derived laboratory strain (NW) were first look at as a standard dataset for SNP discovery. For each rat DNA sample, a final reaction volume of 13.5 µl buffer with 1.5 µl genomic DNA was prepared for PCR. In general, PCR was preformed with an initial denaturation of 94°C (2.5 minutes), followed by 33 cycles of denaturation at 94°C (0.5 minutes), annealing at 57.5°C (40 seconds), and extension at 72°C (1 minute). The final extension was conducted at 72°C (5 minutes). For the primer dimmer cleanup, we added 3.5 µl ExoSAP-IT to 7 µl PCR product, and incubated at 37°C for 15 minutes and then inactivated ExoSAP-IT at 80°C for 15 minutes. PCR product was sent for both forward and reverse sequencing using ABI 3730 xl sequencer (DEWALCH, Houston, TX). Using GenBank sequence NC_005103.2 of *Calu* as the reference gene, SNP sites were identified using Polyphred (Nickerson, Tobe, and Taylor 1997) after contig assembly and editing in Phred/Phrap/Consed (http://www.phrap.org/) (Ewing et al. 1998; Gordon, Abajian, and Green 1998).

With the discovered SNP site Calu_56228735 on *Calu* gene and the site Calu_S1_56038243 on the flanking intergenic region (Table 5.5), we genotyped all the 691 rat samples from 19 natural populations by High Resolution Melting (HRM) test (Wittwer et al. 2003). The HRM PCR Kit was purchased from QIAGEN (http://www.qiagen.com, access Sep 2011). The primers covering the identified SNP sites for HRM genotyping (Appendix 2) were designed in Primer3 (http://frodo.wi.mit.edu/primer3, access May 2011). For each rat DNA sample, 0.65 ul genomic DNA was added to reaction volume of 10 ul, which includes 5 ul HRM PCR Master Mix, 2.95 ul RNase-free water and 1.4 ul primer Mix. The HRM analyses were conducted on the Rotor-Gene Q, with an initial PCR activation of 95°C for 5 min,

followed by 40 cycles of denaturation at 95°C for 10 seconds, annealing and extension at 55°C for 30 seconds; the following HRM were performed in a temperature range from 75°C-90°C with 0.1°C increments at each step. The genotypes for each rat sample were identified with the threshold of confidence percentage > 85% (White, Hall, and Cross 2007) in the HRM analysis of comparison to 6 reference genes with known genotypes.

Genotype-phenotype association test (Table 5.6) between phenotype and SNP sites of *Calu* gene as well as flanking region for 618 rat samples was conducted for each natural and inferred virtual population (based on the population structure analysis) in PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007). Simple Chi-square test of allele model was computed. To control the effect of sex as a potential confounding factor, the Cochran-Mantel-Haenszel (CMH) test was also applied. The phenotype-genotype association of the SNP site Y139C on *Vkorc1* gene was used as a reference.

5.2.7. Linkage disequilibrium between *Vkorc1* and *Calu*

For each natural and inferred virtual population, linkage disequilibrium (LD) among the SNPs in*Vkorc1*, *Calu* gene and Calu_S1 in intergenic reigon were calculated in PLINK (Purcell et al. 2007) using the measure of squared r, which represents the correlation between two loci (Table 5.7).

5.3. Results

Inferring population structure is a critical step before performing association analyses. Otherwise, spurious associations would be detected, which might be due to the population admixture or demographic history (Mezmouk et al. 2011). We first collected genotype data of microsatellites and presumably neutral SNPs for ~700 rats. The genotypes were called based on high resolution melting (HRM) genotyping data using two combined clustering methods in R. Based on the genotype data of neutral sites, we examined the population structure for 618 wild rats of 19 farms from a resistant area in Germany (Figure 5.1A) (Kohn, Pelz, and Wayne 2000; Kohn, Pelz, and Wayne 2003). Samples were assigned to inferred virtual populations based on the STRUCTURE and INSTRUCT population membership assignment. Then considering both the natural population and virtual population structure, we selected 46 wild rats for SNP array experiment and performed the genome wide association study to detect and validate candidate genes involved in warfarin resistance.

5.3.1. Population structure analysis inferred 3 virtual populations with genetically homogeneous individuals

We implemented Bayesian method in STRUCTURE to cluster individuals by minimizing the within cluster linkage disequilibrium and keeping Hardy-Weinberg equilibrium (Pritchard, Stephens, and Donnelly 2000). Besides, we also performed the analysis in INSTRUCT allowing inbreeding model and without the assumption of Hardy-Weinberg equilibrium. A network clustering approach (NetView) was recently developed

to detect population structure; its performance needs more evaluation (Neuditschko, Khatkar, and Raadsma 2010).

First, we ran the cluster number from 1 to 19 to choose the optimal cluster value K. Determining the optimal number of clusters of genetically homogeneous individuals is always a challenge. Here we used 7 different criterions to detect K: 1) the posterior probability of the data Ln Pr(X|K)for each given K; 2) the second order rate of change of Ln Pr(X|K) – delta K; 3) the average maximum correlation coefficients; 4) the Neighbor-Joining tree; 5) the deviance information criterion (DIC) for INSTRUCT results; 6) the second order rate of change of F_{ST} – delta F_{ST} for STRUCTURE results; 7) the value of dirichlet parameter alpha for STRUCTURE results. See Appendix 9 for detailed information about the support of K for each measure. We collected these different measures (to the best of our knowledge) to detect K, which would be a useful guide for further and other population structure studies.

Overall (Table 5.1), 5 measures suggested optimal K=3, and 6 measures supported the optimal K = 6. From the NJ tree, we see only with K=6, the lines were clearly clustered (Appendix 9). This situation implies K=3 is suitable and there might be hierarchical structures. For following analysis, we chose K=3.

Table 5.1 – Choosing optimal K in population structure analysis

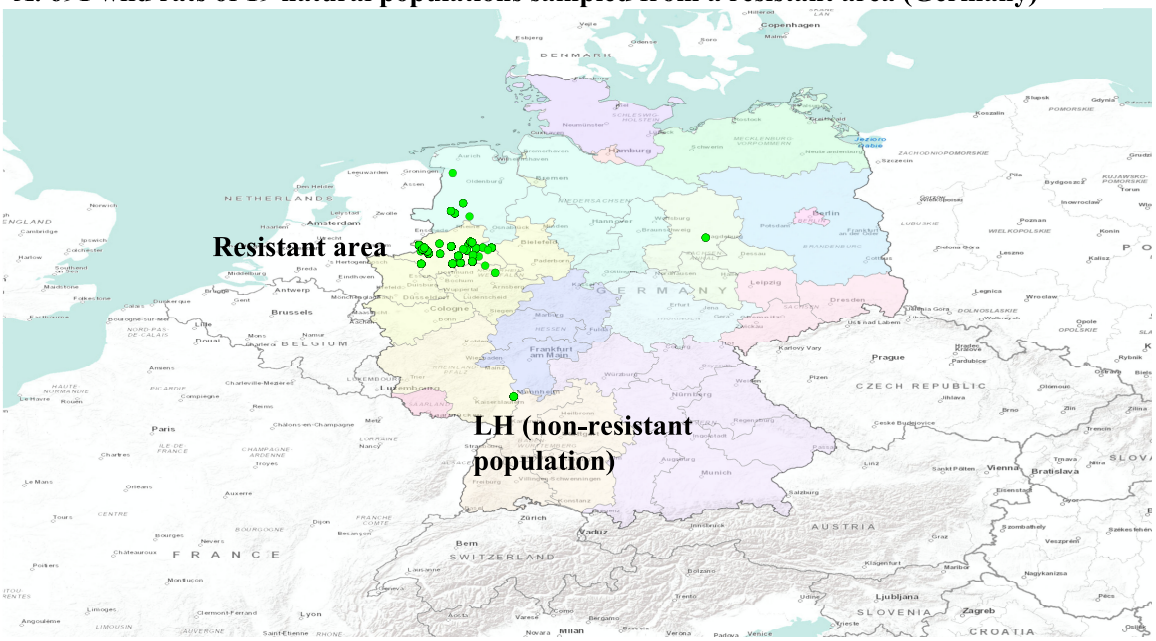
Criteria	STRUCTURE	INSTRUCT
LnPr(X|K)	6	6
ΔK	3	3
AverMaxCorr	6	3
NJ tree	6	NA
DIC	NA	6

ΔF_{ST}	3	3
alpha	6	NA

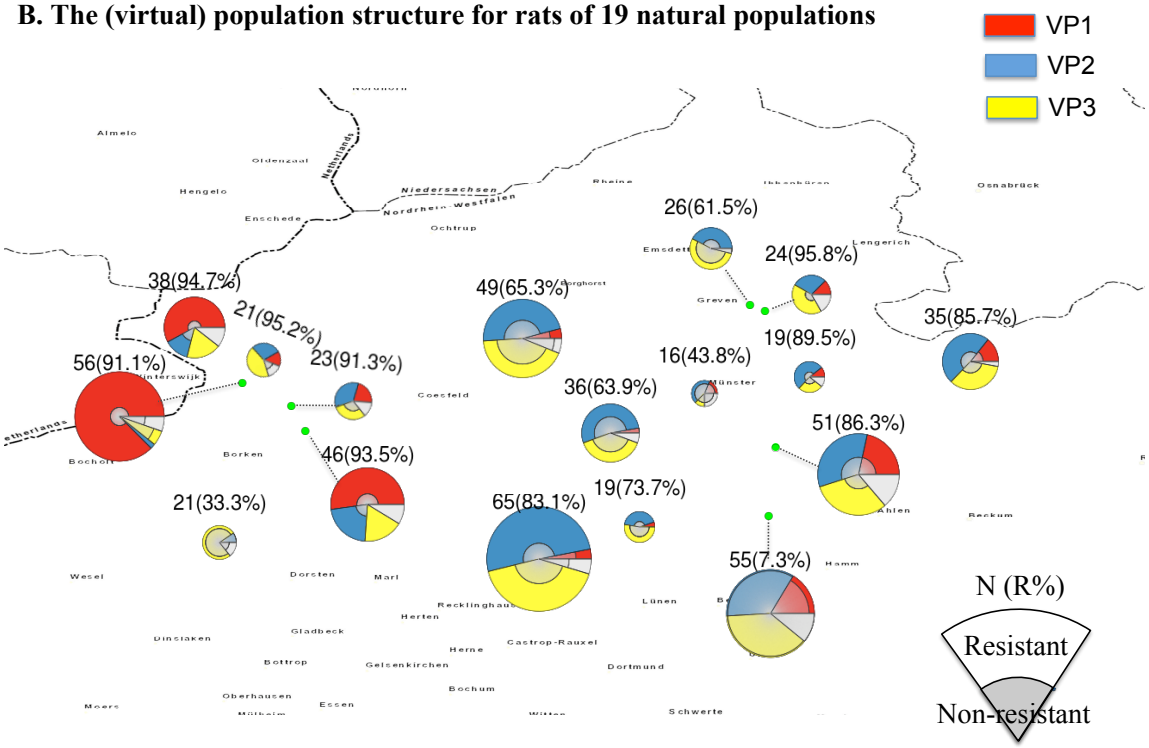
See Appendix 9 for details about choosing K based on different criteria.

Figure 5.1B demonstrated the population structure and the statistics for each natural population in the resistant area in Germany. We chose 46 rat samples from 3 natural populations and 3 virtual populations. As shown in Figure 5.1C, VP1 is composed of samples from two natural populations: WU and TH; whereas VP2 and VP3 contain samples all from population 20. Then in this case, we could perform GWAS in either natural populations or virtual populations.

A. 691 wild rats of 19 natural populations sampled from a resistant area (Germany)



B. The (virtual) population structure for rats of 19 natural populations



C. 46 samples from 3 natural and 3 virtual populations for SNP array II

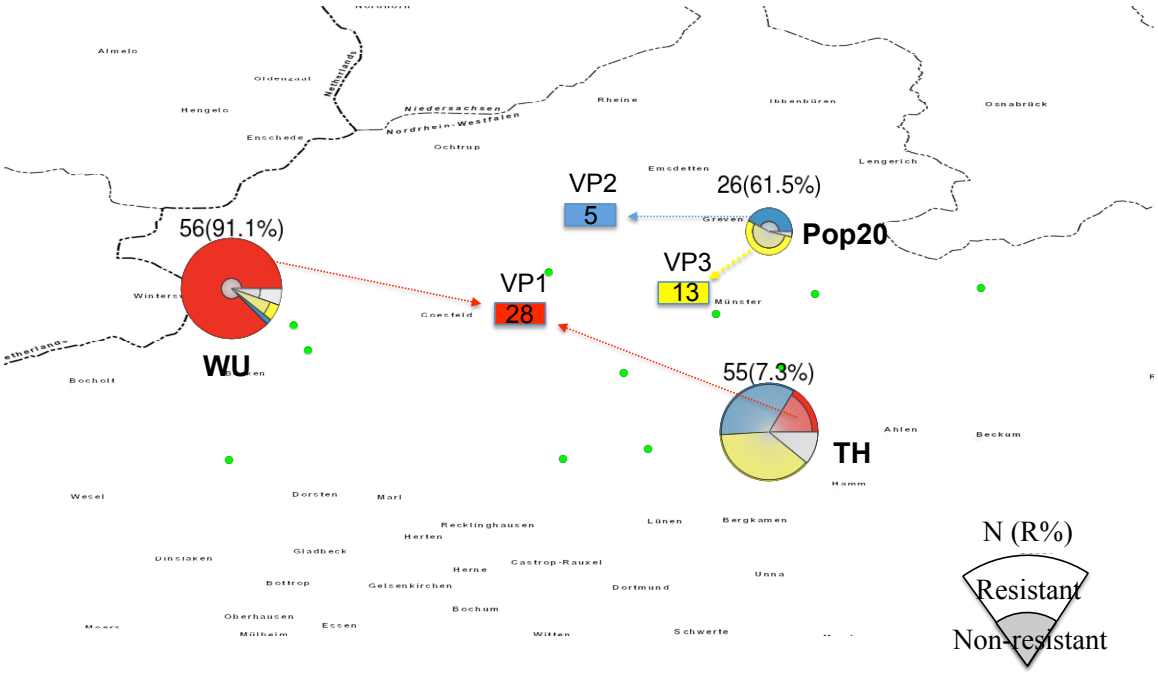


Figure 5.1 – Maps of wild rats with population structure information. A. Wild rats were sampled from a resistant area at northwestern Germany. Non-resistant rats in LH population were sampled 300 km away. B. The inferred genetically homogeneous population structure for 19 natural populations. C. The population structure for 46 samples selected for SNP array II experiment.

5.3.2. GWAS identified 8 top candidate SNPs in natural populations

We didn't pool samples from SNP array I and SNP array II together for the genomic association test because of the high inflation factor ($\lambda = 2.64$). Even within SNP array II, there was high inflation factor ($\lambda = 2.33$) indicating the existence of population structure, which is expected since SNP array II is composed of samples from 3 natural populations. Considering population structure and sample size, for these SNP array II samples, we tested genomic association in two groups: virtual population 3.1 (28 samples, VP3.1 composed of WU and TH samples) and pop20 (5 VP3.2 and 13 VP3.3 samples). The inflation factor in VP3.1 is still high ($\lambda = 2.13$), but in pop20 is relatively low ($\lambda = 1.26$). Nevertheless, these inflation factors suggest the necessity of genomic correction.

Vkorc1 gene is still significantly associated with warfarin resistance phenotype in both VP3.1 (P-value = 1.7×10^{-4}) and pop20 (P-value = 4.3×10^{-3}), but surprisingly not after corrected for multiple tests (HOLM P-value = 1). There are 54 SNPs in VP3.1 with significance level smaller than *Vkorc1*'s P-value and 8 SNPs are significantly associated with resistance phenotype after multiple correction (P-value < 0.05) (Table 5.2). In pop20, there are 23 SNPs with significance level smaller than *Vkorc1*'s, but none of them

are still significant after multiple correction. There is no overlap of the 54 SNPs in VP3.1 and the 23 SNPs in pop20.

Comparing with the 82 top associated SNPs from SNP array I (c.f. Chapter 3), there is one SNP (S425367) at chromosome 16 (position: 4663834 bp) detected as top SNP based on both SNP array I and array II data. There are two genes within 0.5 Mb distance to this SNP: *LOC100360666* and *Cacna2d3* (calcium channel, voltage-dependent, alpha 2/delta subunit 3). Interestingly, another calcium-channel gene *CACNA1C* (calcium channel, voltage-dependent, L type, alpha 1C subunit) was noticed previously in a human genomic scan for warfarin dose associated genetic variants (Cooper et al. 2008). The rat orthologs *Cacna1c* gene and *Cacna2d3* gene are directly interacting according to STRING database (http://string-db.org). And one of the 8 SNPs significantly associated with resistance after multiple correction, SNP (S423237, Table 5.2) at chromosome 4 (position: 14219587 bp) is also near a calcium-channel gene *Cacna2d1* (calcium channel, voltage-dependent, alpha2/delta subunit 1). Another SNP (S422784) among these 8 SNPs is located in a gene *Fhit* (fragile histidine triad), which has been identified as a candidate gene in NetGWAS based on SNP array I (Appendix 4).

Table 5.2 – Top associated SNPs from GWAS based on SNP array II data

SNP	Chromosome	Position (Mb)	P-value	HOLM_Pvalue
S422784	15	17	3.55e-7	0.002
S423237	4	14	3.94e-7	0.003
S697166	1	105	5.75e-7	0.004
S693878	1	152	1.61e-6	0.010
S423282	1	86	2.74e-6	0.018
S693984	3	22	3.59e-6	0.023

S692698	4	60	5.39e-6	0.035
S695878	15	102	5.44e-6	0.035

P-value: uncorrected P-value of genomic association test.
HOLM Pvalue: P-value after multiple correction using HOLM method.

5.3.3. NetGWAS identified *Vkorc1* and other candidate genes

After applying the modified PageRank algorithm to the SNP array II data, we obtained 159 candidate genes and 143 candidate genes from VP3.1 and pop20 respectively. Looking at these candidate genes, we found some of them are from the same chromosomal region. So if we group these candidate genes into a cluster when there are more 4 genes within 2 Mb distance from each other, we obtained 13 clusters in VP3.1 and 12 clusters in pop20. A region at chromosome 7 (114-116 Mb) containing 27 candidate genes detected in VP3.1 gained our attention. There is an *Apol3* gene (apolipoprotein L, 3) maybe functional related. As in human, vitamin K_1 is taken along with the dietary fat, which is then cleared by *APOE* gene (apolipoprotein E) (Wadelius et al. 2007).

If applying the same grouping criteria to previous SNP array I results, we got 7 clusters. And there are 21 regions of clustered genes from the microarray analysis (Table 4.2). We compared these regions with above gene clusters in VP3.1 and pop20. The regions of gene clusters identified from both microarray and VP3.1 or pop20 might be relatively important because they are replicates from different populations (Table 5.3). In addition, there are 3 regions (Chr1: 184-187, Chr12: 32-37 and Chr20: 1-5 Mb) repeatedly appeared. Obviously *Vkorc1* is located in the first region. In chromosome 12: 32-37 Mb region, there is a *Tbx3* (T-box 3) gene, which is involved in reproductive developmental

process. Interestingly, it might be indirectly interacting with the vitamin K dependent protein - bone matrix protein MGP because *Tbx3* is mediated by bone morphogenetic protein (BMP) signals and MGP modulates BMP-2 activity (Zebboudj, Shin, and Bostrom 2003; Chen et al. 2009).

We also specifically compared these candidate genes with previously identified 87 candidate genes from SNP array I and candidate targets from microarray (Table 5.3). We saw *Vkorc1* is shown in all populations as expected. *Ggcx* gene was identified from both SNP array I and VP3.1, which is involved in vitamin K pathway to activate vitamin K dependent proteins. But previous tests of its association with warfarin dose in human reported controversial results (Wadelius et al. 2005; Cha et al. 2007; Wadelius et al. 2007), so it is our chance to test the *Ggcx* gene in rat populations. Gene *Fhit* might be of interest as it has been identified as candidate genes based on SNP array I and II data, plus the SNP (S422784) in it is significantly associated with resistance phenotype (Table 5.2). Other gene names are not familiar in terms of known vitamin K pathway genes, but might be useful for future examination.

Table 5.3 - Comparison of candidate gene list between SNP array I and II and microarray

Comparison	**Candidate genes**
SNPArrayI vs. Microarray	***Vkorc1****, Bglap, Ift52, Mybl2, Cyp4f1, Tbx3, Mlycd, Osgin1, RT1-S3*
SNPArrayI vs. VP3.1	***Vkorc1****, Tuft1, Ggcx, Rock2, Fhit, Tpmt*
SNPArrayI vs. pop20	***Vkorc1***
VP3.1 vs. pop20	*Gsg1l, LOC100360815,* ***Vkorc1****, Hdc, Myh9, Megf11, Rab11a, Tbc1d5, Ncam2, Abcc5, Clph, Csn3, LOC494224, Odam, Csn1s2b, Csn1s2a*
Comparison	**Regions of clustered candidate genes (Mb)**

SNPArrayI vs. Microaray	**Chr1: 185-187**; Chr2: 178-182; Chr3: 153-158; Chr4: 9-13; **Chr12: 33-38**; Chr19: 49-53; **Chr20: 1- 3**
Microarray vs.VP3.1	Chr1: 82-86; **Chr1: 184-187**; Chr7: 113-116
Microarray vs. pop20	**Chr1: 184-187**; Chr11: 82-83; **Chr12: 32-37**; **Chr20: 1-5**
VP3.1 vs. pop20	**Chr1: 185-187**; Chr14: 21

5.3.4. Summary of candidate genes combining NetGWAS of two SNP array data and gene expression analysis

Combining the results of SNP array I, SNP array II and microarray analysis, we listed some candidate genes for further validation (Table 5.4). *Vkorc1* gene has been reconfirmed in three populations by association tests. In microarray analysis, although the adaptive mutation in *Vkorc1* changes protein structure rather than modifying expression profiles, the genes in flanking region exhibited warfarin-related expression signals (Figure 4.6A) due to the effect of selective sweep (c.f. Chapter 2). Thus *Vkorc1* could be detected based on both variation and expression signals.

Cyp4f1 gene gains support from both SNP array I and microarray analysis. It is one of the ortholog gene of human *CYP4F2* gene, which has been identified as a warfarin dose predictor explaining ~1.5% dose variance in patients (Pautas et al. 2009; Takeuchi et al. 2009). In vitro metabolic experiments showed that *CYP4F2* gene is a vitamin K_1 oxidase, and the V433M allele in *CYP4F2* result in a reduced enzyme activity, thus an elevated level of hepatic vitamin K, which requires more warfarin dose to elicit the same anticoagulant response (McDonald et al. 2009). Phylogenetic analysis also indicated that CYP4F genes are subject to positive selection (Thomas 2007). These information suggest

Cyp4f1 gene might be involved in warfarin resistance in rats, thus would be of interest for further validation. However, as recognized before, *CYP450* genes duplicate a lot and enzymes function in xenobiotic substrates tend to form clusters (Thomas 2007). Based on GenBank annotation, we saw that closely near *Cyp4f1* gene is a *Cyp4f6* gene. And ~0.3 Mb away, there are two genes *Cyp4f40* and *Cyp4f4*, which is actually another rat ortholog of *CYP4F2* gene. Expanding to 0.6 mb, there are 4 more *CYP450* genes. This situation and the rapid evolutionary rate of xenobiotic related *CYP450* genes pose the challenge of discerning the 'true' warfarin related gene, because it is difficult to assign one-to-one ortholog for these genes.

Table 5.4 – Summary of candidate genes or regions

Genes or Regions (Mb)	Chr	Pos (Mb)	Human ortholog	Function	SNP arrayI (NW)	Micro array (NW)	SNP arrayII (VP3.1)	SNP arrayII (pop20)
82-86	1	82-86	-	-		✓	✓	
Vkorc1	1	187	*VKORC1*	regulation of blood coagulation	✓	✓	✓	✓
Fgfr2	1	189	*FGFR2*	bone development, *FGFBP2*	✓	✓		
Bglap	2	180	*BGLAP*	bone mineralization	✓	✓		
Ift52 (153-158)	3	154	*IFT52*	intraflagellar transport 52 homolog	✓	✓		
Mybl2 (153-158)	3	154	*MYBL2*	myeloblastosis oncogene-like 2	✓	✓		
Calu	4	56	*CALU*	regulate vitamin K cycle	✓			✓
Ggcx	4	106	*GGCX*	gamma-carboxylation	✓		✓	
Cyp4f1	7	14	*CYP4F2*	cytochrome P 450 gene family	✓	✓		
113-116	7	113-116	-	*Apol3* apolipoprotein, *APOE*		✓	✓	
82-83	11	82-83	-	-		✓		✓
Tbx3 (32-37)	12	38	*TBX3*	reproductive development	✓	✓		✓
21-21	14	21	-	-			✓	✓

Fhit	15	16	*FHIT*		✓		✓	
Cacna2d3	16	4	*CACNA2D 3*	calcium channel protein, *CACNA1C*	✓		✓	
Mlycd (49-53)	19	50	*MLYCD*	fatty acid metabolic process	✓	✓		
Osgin1 (49-53)	19	50	*OSGIN1*	growth factor activity	✓	✓		
RT1 (1-5)	20	2	*HLA-E*	histocompatibility complex gene	✓	✓		✓

Bglap gene, encoding a vitamin K dependent protein that affects vessel calcification and bone mineralization (Suttie 1993; Danziger 2008), is ascertained from both SNP array and microarray study. More interestingly, our preliminary test of genomic association with arterial calcification phenotype indicated that 4 SNPs near *Bglap* gene are significant associated with both arterial calcification and warfarin resistance phenotypes (c.f. Chapter 6, Future directions). Considered previously observed arterial calcification of homozygote resistant rats (Kohn, Price, and Pelz 2008), *Bglap* gene might open a door for us to explore the mechanisms of one of the fitness costs associated with warfarin resistance besides the reduced growth and reproductive rate (Jacob et al. 2012). The other two vitamin K dependent proteins (*Gas-6* and *Mgp*) with the function of protecting vasculature, however, were not detected in our study (Danziger 2008). In vitro experiment showed that MGP's effect on calcification also relies on its relative amounts with the bone morphogenetic protein BMP-2 (Zebboudj, Shin, and Bostrom 2003). In this complicate scenario, it is hard to predict what we would observe for *Mgp* gene.

The region of 154-158 Mb on chromosome 3 exhibited relatively strong association signals as well as expression signals. However, there are 9 genes in this region; which one would be the potential causal gene? At first, the *Prex1* gene carried the

strongest associated SNP seemed to be an interesting one. I have conducted SNP discovery and genotyping for this gene in other wild populations, but didn't find significant association with warfarin resistance. Based on the gene interaction network analysis (Figure 3.7A), the *Prex1* gene was not connected to the main *Vkorc1* cluster in the GGI subnetwork of candidate genes. Instead, other two genes (*Ift52* and *Mybl2*) in this region carried relatively high association scores and also connected to *Vkorc1* cluster in terms of gene interaction, thus might be worthy of further examination. Moreover, genes in this region interact with the *Fgfr2* gene (involved in embryonic development and tissue repair) in *Vkorc1* sweep region, which suggests us another path to explore the association with warfarin resistance for this region.

There are 3 other genes *Tbx3*, *Mlycd* and *RT1-S3*, with supports from expression and genomic variation analysis. Especially *Tbx3* gene was identified from two SNP array and the microarray data. We suspect that *Tbx3* gene might be indirectly interacting with the vitamin K dependent protein - bone matrix protein (MGP) via the bone morphogenetic protein (BMP-2) (Zebboudj, Shin, and Bostrom 2003; Chen et al. 2009). And the functional enrichment analysis (Table 3.2) showed that *Tbx3* gene is involved in reproductive developmental process, which is an interesting coincidence that resistant rats were observed to have fitness costs of both arterial calcification and reduced reproductive rate (Smith, Townsend, and Smith 1991; Pelz et al. 2005; Kohn, Price, and Pelz 2008; Rost et al. 2009). Though lack of enough knowledge of their relationship with warfarin, these genes should be paid attention to in future analysis.

As we mentioned above, *Ggcx*, *Fhit* and *Cacna2d3* genes were identified based on SNP array data from two populations. *Ggcx* gene were examined as a candidate gene related to warfarin before but was not fully confirmed in replication studies (Wadelius et al. 2005; Cha et al. 2007; Wadelius et al. 2007). *Cacna2d3* is potentially interesting because there is another calcium-channel gene *CACNA1C* carried the best novel SNP in a GWAS of warfarin dose in human (Cooper et al. 2008). Another gene *Fgfr2* (fibroblast growth factor receptor 2), shown to be involved in bone development (Table 3.2), also have supporting evidence from human: the genomic scan revealed *FGFBP2* (fibroblast growth factor binding protein 2) gene has a relatively high genotype-dose association in addition to *VKORC1* and *CYP2C9* in human (Cooper et al. 2008).

There are two regions are supported by microarray and SNP array II data in VP3.1. One region is at chromosome 1: 82-86 Mb. Another region is at chromosome 7: 114-116 Mb, containing 27 candidate genes. There is an *Apol3* gene (apolipoprotein L, 3) in this region. And in human, another *APOE* gene (apolipoprotein E) is responsible for the transportation of vitamin K_1 from upper gastrointestinal tract to liver. In addition, there is a region at chromosome 14 around 21 Mb carrying replicate signals from VP3.1 and pop20. 6 genes in this region need more literature search for function relevance.

According to the traditional candidate approach, it is intuitive to examine the rat ortholog of the human *CYP2C9* gene, which is the secondary warfarin dose biomarker just behind *VKORC1*. Human *CYP2C9* gene and its rat ortholog *Cyp2c11* were found under positive selection based on phylogenetic analysis (Thomas 2007). I did a small-scale SNP discovery for rat *Cyp2c11* gene, but didn't find significant association with

warfarin resistance. More replication experiments are needed to assess the role of *Cyp2c11* gene in rat resistance.

5.3.5. Candidate gene *Calu* was associated with warfarin resistance in multiple wild populations

As we are listing these top candidate genes for further evaluation, we actually already collected genotype data of another candidate gene *Calu* in multiple wild populations. Calumenin gene (*Calu*), located at chromosome 4: 56 Mb, has been identified as a candidate in SNP array I analysis (c.f. Chapter 3) with relatively high support score, but without high association strength in the wild-derived NW populations. In pop20 (SNP array II), three genes located at the same chromosomal region were detected as candidate genes. As previous studies showed that *Calu* regulate the vitamin K-dependent gamma-carboxylation system (Markussen et al. 2007b) and compete with warfarin for the binding-site in the VKOR complex (Markussen et al. 2007a), we performed SNP discovery for *Calu* gene and collected genotype data.

Using 8 primer pairs covering 7 exon regions on the *Calu* gene, we found 4 SNPs located on 5' region, 1 SNP on the first intron, 1 SNP, a 26bp deletion and a 2bp insertion on the 5th intron, 3 SNPs on the 6th exon (Table 5.5). For the SNP site Calu_56228735 located on the first intron, we collected the genotype information using HRM for all the rats from the 19 natural populations and used it in the association test. The intergenic SNP site Calu_S1_56038243 was also genotyped as a linked locus to *Calu* gene (0.2 upstream).

Table 5.5 – Characterization of SNP variants in *Calu* gene and flanking region

Genes	Position	SNP ID	Location	Variation
Vkorc1	Chr1:187177048	**Vkorc1_187177048**	*Vkorc1* 1st exon	A->G
Calu 5' region	Chr4:56038243	**Calu_S1_56038243**	*Calu* upstream intergenic	C->G
Calu	Chr4:56228407	Calu_56228407	*Calu* 5' intergenic	C->G
Calu	Chr4:56228527	Calu_56228527	*Calu* 5' intergenic	G->T
Calu	Chr4:56228578	Calu_56228578	*Calu* 5' intergenic	C->T
Calu	Chr4:56228735	**Calu_56228735**	*Calu* 1st intron	A->G
Calu	Chr4:56251705	Calu_56251705	*Calu* 5th intron	C->A
Calu	Chr4:56251821	Calu_56251821	*Calu* 5th intron	26bp deletion
Calu	Chr4:56252081	Calu_56252081	*Calu* 5th intron	2bp insertion
Calu	Chr4:56252174	Calu_56252174	*Calu* 6th exon	1->2
Calu	Chr4:56252656	Calu_56252656	*Calu* 6th exon	G->A
Calu	Chr4:56252715	Calu_56252715	*Calu* 6th exon	G->A

Bold SNP IDs are the SNPs used in the following association tests.

As the main resistance gene, *Vkorc1* showed significant associations ($P < 0.031$) with the resistant phenotype in NW and 13 natural populations (Table 5.6). Represented by the SNP Calu_56228735, *Calu* gene exhibited significant association ($P < 0.02$) in 7 populations and marginal association ($P = 0.08$) in NW population. The intergenic SNP site Calu_S1_56038243 (call it Calu_S1 from now on) located 0.2 Mb upstream to the *Calu* gene also showed significant associations ($P <= 0.017$) in 7 populations. This observation indicates *Calu* is associated with warfarin resistance and the flanking region was also affected by warfarin selection due to potential hitchhiking effect.

There are 5 populations missing *Vkorc1* association signals. In Pop6, Pop10 and Pop25, there are too few susceptible rats (≤ 2); in pop15, only 1 rat out of 6 is resistant. Besides the small number of non-resistant rats in above 3 populations, these non-resistant rats actually carry the Y139C mutation in *Vkorc1* gene. It might be due to some misclassification of BCR (blood clotting response) test, or implies the effect of other

genes on warfarin resistance. In the SP population with moderate resistant level (33%), we noticed the absence of phenotype-genotype association for *Vkorc1* gene, which can hardly be explained by sample size (7 resistant and 14 non-resistant rats). In SP population, all the samples (except one sample 3452), even the 6 resistant rats, are homozygotes of the *Vkorc1* non-resistance allele A. This observation suggests other genes may contribute to wafarin resistance, and is one motivation for us to do genomic scan for additional warfarin related genes. And here we detected significant association in this SP population at the Calu_S1 locus, which merits further investigation. The significant association for both *Vkorc1* and *Calu* gene are maintained in the inferred virtual populations.

Table 5.6 – Warfarin resistance association tests for *Calu* in multiple wild populations.

Pop	N	R(N)	S(N)	R%	*Vkorc1*	*Calu_S1*	*Calu*
NW	32	20	10	69%	2.6e-7*	0.110	0.080
WU	73	51	4	93%	9.0e-5*	0.000*	0.017*
KB	70	54	11	83%	1.5e-4*	0.465	0.780
SP	23	7	14	33%	0.273	0.002*	0.251
TH	56	4	51	7%	0.001*	0.932	0.768
Pop4	52	43	3	94%	0.008*	0.252	0.008*
Pop17	53	44	7	86%	1.4e-6*	0.010*	0.756
Pop28	37	30	5	86%	0.028*	0.011*	0.015*
Pop5	20	14	5	74%	0.004*	0.767	0.020*
Pop19	56	32	17	65%	3.1e-8*	0.000*	0.011*
Pop13	42	23	13	64%	1.5e-5*	0.257	0.266
Pop20	27	16	10	62%	0.000*	0.017*	0.012*
Pop14	19	7	9	44%	0.001*	0.007*	0.008*
Pop15	10	1	5	17%	0.083	0.693	0.127
Pop6	24	23	1	96%	0.103	0.711	0.106
Pop12	25	20	1	95%	0.012*	0.307	0.763
Pop10	42	36	2	95%	0.170	0.696	0.419
Pop18	29	21	2	91%	0.031*	0.200	0.103
Pop25	20	17	2	90%	0.278	0.071	0.156
Virtual_pop1	158	122	20	77%	1.8e-10*	1.6e-3*	7.5e-7*
Virtual _pop2	207	154	51	74%	2.5e-25*	6.7e-5*	4.7e-3*
Virtual _pop3	210	129	77	61%	8.4e-34*	3.8e-9*	1.1e-8*

N: Sample size for each population.
R(N) and S(N): the number of rat samples with the warfarin resistant (R) and susceptible (S) phenotype.
R%: the warfarin resistant level, the number of resistant rats divided by the number of samples with warfarin phenotype.
* are the significant P-values < the threshold of 0.05.

5.3.6. *Vkorc1* and *Calu* were in linkage disequilibrium

Linkage disequilibrium (LD) described the non-random association of alleles between different SNP sites (Hartl and Clark 2007). For SNPs located on the same chromosome, LD would decrease with the increase of distance (Xiong and Guo 1997). Simultaneously, LD is also affected by recombination rate and population structure (Xiong and Guo 1997). Extended LD blocks of linked loci are usually considered to be signatures of selection (Chapter 2, Figure 2.2A) (Kohn, Pelz, and Wayne 2000; Kim and Nielsen 2004; Hohenlohe et al. 2011). Thus, we expected to detect strong LD signals at regions with potential adaptive variants in populations resistant to warfarin. Sometimes, however, LD might be maintained by functional interaction between SNP pairs even located on different chromosomes; reciprocally, strong LD between physically unlinked sites with biological significance would help identify functional relationship between the SNP pairs (Burmeister 1999; Slatkin 2008).

Table 5.7 showed the LD between SNPs with the threshold of squared r >= 0.15. SNPs in the *Calu* gene are in LD with its upstream intergenic SNP Calu_S1, indicating the effect of selective sweep. Though distributed on different chromosomes, *Vkorc1* is in LD with the *Calu* gene as well as its upstream intergenic SNP in multiple wild populations. These non-random associations between physically unlinked genes implied

their potential functional relatedness or they were under the same selection pressure (Liu et al. 2010).

Table 5.7 – Linkage disequilibrium of *Vkorc1* and *Calu* in multiple wild populations.

Population	SNPA	Position_SNPA	SNPB	Position_SNPB	LD (r-square)
NW	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.16
NW	Calu_S1	Chr4: 56038243	Calu	Chr4: 56228736	0.48
WU	Calu_S1	Chr4: 56038243	Calu	Chr4: 56228736	0.16
Pop19	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.15
Pop19	Vkorc1	Chr1: 187177049	Calu_S1	Chr4: 56038243	0.19
Pop4	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.11
Pop28	Vkorc1	Chr1: 187177049	Calu_S1	Chr4: 56038243	0.13
Pop20	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.28
Pop12	Vkorc1	Chr1: 187177049	Calu_S1	Chr4: 56038243	0.11
Pop12	Calu_S1	Chr4: 56038243	Calu	Chr4: 56228736	0.27
Pop5	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.18
Pop25	Vkorc1	Chr1: 187177049	Calu_S1	Chr4: 56038243	0.21
Pop25	Calu_S1	Chr4: 56038243	Calu	Chr4: 56228736	0.18
Pop14	Vkorc1	Chr1: 187177049	Calu_S1	Chr4: 56038243	0.41
Pop14	Vkorc1	Chr1: 187177049	Calu	Chr4: 56228736	0.32
Pop14	Calu_S1	Chr4: 56038243	Calu	Chr4: 56228736	0.11

5.4. Discussion

Based on genomic variation and expression data, we performed network-guided analysis and obtained a list of candidate genes related to warfarin. Most of them are involved in or related to vitamin K pathway, and are connected to the resistance gene *Vkorc1* in the gene-gene interaction network (Figure 3.7). Here we reviewed previously examined candidate genes considering the warfarin therapy in human, which are important reference for our and future study.

5.4.1. Short review of 31 previously studied candidate genes in human

As the most popular oral anticoagulant for preventing thrombotic events, warfarin therapy is associated with risk of thrombosis or risk of bleeding, especially during the initial phase (Pavani et al. 2012).

In human, *VKORC1* and *CYP2C9* established the warfarin dose prediction model coupled with other non-genetic factors (sex, age, weight, etc.). *CYP4F2* and *EPHX1* genes are recently identified with small influence on warfarin dose (Pautas et al. 2009; Takeuchi et al. 2009). As shown in Figure 5.2, *VKORC1* encode vitamin K epoxide reductase; -1639 G to A mutation (rs9923231) in the upstream promoter region produces less VKORC1, thus lower warfarin dose is needed for patients with the A allele. More nonsynonymous mutations in *VKORC1* associated with warfarin sensitivity have been detected recently and complicate the dose optimization algorithms (Baniasadi et al. 2011; Pavani et al. 2012). *CYP2C9* and *CYP4F2* encode enzymes metabolizing S-warfarin and vitamin K_1 respectively (McDonald et al. 2009). Patients with the variant alleles *2 (R144C, rs1799853) and *3 (I359L, rs1057910) in *CYP2C9* gene have greater risks of bleeding and need lower warfarin dose (Takeuchi et al. 2009). Whereas individuals with the V433M (rs2108622) allele in *CYP4F2* gene require higher warfarin dose because the allele reduce enzyme capacity to metabolize vitamin K_1 (McDonald et al. 2009).

Though there remains ~40% unexplained variance of warfarin dose, neither candidate gene approach nor genomic scan identified new determinants of warfarin dosing (Wadelius et al. 2007; Takeuchi et al. 2009). About 30 candidate genes believed to be involved in vitamin K or warfarin interactive pathways were examined: *PROC*,

APOE, *EPHX1*, *ORM1-ORM2*, *CALU* and *GGCX* were associated with warfarin dose, but none except *VKORC1* and *CYP2C9* reach significance after multiple correction (Wadelius et al. 2007).

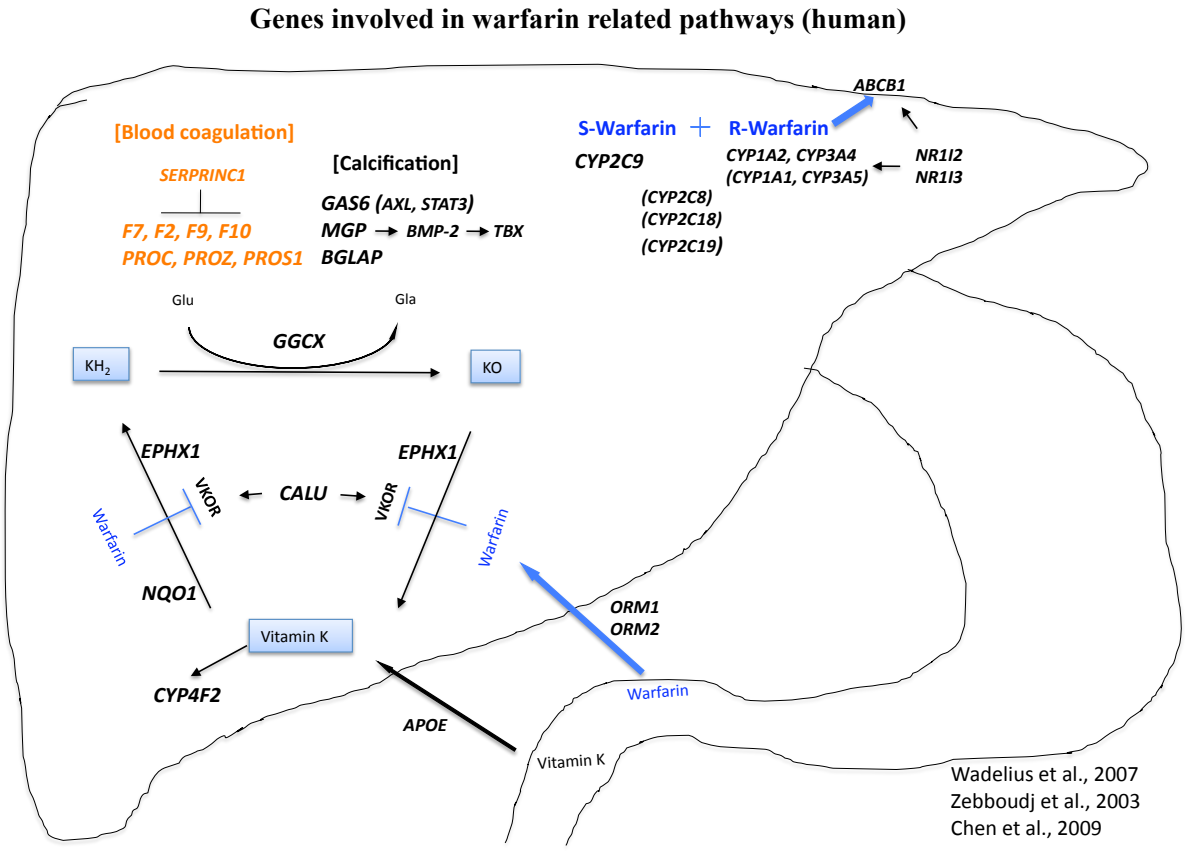


Figure 5.2 – Overview of the interaction between warfarin and genes involved in vitamin K related pathway based on human studies.

For future reference, we briefly introduced the genes involved in warfarin interactive pathways base on research in human (Wadelius et al. 2007) (Figure 5.2, Appendix 8). Warfarin inhibited VKOR (encoded by *VKORC1*) reductase activity, thus

impaired the recycling of vitamin K hydroquinone, which is an essential cofactor for the formation of functional vitamin K dependent proteins (mostly blood coagulation factors) by gamma-carboxylase, encoded by *GGCX* (Presnell and Stafford 2002; Pelz et al. 2005; Stafford 2005). Another Vitamin K epoxide reductase is encoded by *EPHX1* gene, which was found to be associated with the maintenance dose of warfarin (Pautas et al. 2009). Calumenin, encoded by *CALU* gene, has been shown to regulate the vitamin K-dependent gamma-carboxylation system (Markussen et al. 2007b) and compete with warfarin for the binding-site in the VKOR complex (Markussen et al. 2007a). *APOE* (apolipoprotein E) and *NQO1* (NAD(P)H dehydrogenase, quinone 1) gene encode apolipoprotein E and NAD(P)H dehydrogenase, which the potential to affect dietary vitamin K (Markussen et al. 2007b; Wadelius et al. 2007).

As mentioned above, several vitamin K dependent proteins are converted into biologically active forms by the gamma-carboxylation process, which requies vitamin K hydroquinone as a cofactor. These genes are involved in blood coagulation and bone metabolism: *F2* (coagulation factor II), *F7* (coagulation factor VII), *F9* (coagulation factor IX), *F10* (coagulation factor X), *PROC* (protein C, inactivator of coagulation factors Va and VIIIa), *PROS1* (protein S (alpha)), *PROZ* (protein Z, vitamin K-dependent plasma glycoprotein) and *MGP* (matrix Gla protein), *BGLAP* (bone gamma-carboxyglutamate (gla) protein) (Suttie 1993; Stafford 2005; Danziger 2008). *SERPINC1* (serpin peptidase inhibitor, clade C (antithrombin), member 1) gene encoded a non-vitamin K-dependent protein, which inhibits the Vitamin K dependent clotting factors (Wadelius et al. 2007).

Other vitamin K dependent proteins are involved in calcification process. The growth-arrest specific protein 6, encoded by *GAS6*, affects vascular smooth muscle cell movement and apoptosis (Yanagita 2004; Danziger 2008). MGP (bone matrix gla protein) is a vascular calcification inhibitor, interacting with bone morphogenetic protein BMP-2 (encoded by *BMP2* gene, bone morphogenetic protein 2), which further mediate *TBX genes* (Zebboudj, Shin, and Bostrom 2003; Chen et al. 2009). *BGLAP* is also related to vessel calcification. Different from coagulant factors that are produced and carboxylated in liver, MGP and GAS6 are produced and carboxylated locally within vasculature (Danziger 2008).

Cytochrome P450 enzymes metabolize warfarin in liver. *CYP2C8* (cytochrome P450, family 2, subfamily C, polypeptide 8), especially *CYP2C9* play important role in metabolizing the S-form of administered warfarin. *CYP1A2* (cytochrome P450, family 1, subfamily A, polypeptide 2), *CYP3A4* (cytochrome P450, family 3, subfamily A, polypeptide 4), *CYP1A1* (cytochrome P450, family 1, subfamily A, polypeptide 1), *CYP3A5* (cytochrome P450, family 3, subfamily A, polypeptide 5) were reported to metabolize the R-warfarin. An interestingly study on habitual caffeine consumption demonstrated that *CYP1A2* gene was the primary enzyme involved in caffeine metabolism (Cornelis et al. 2011). Other genes like *ORM1* (orosomucoid 1), *ORM2* (orosomucoid 2) and *ABCB1* (ATP-binding cassette, sub-family B (MDR/TAP), member 1) encoded proteins involved in warfarin transportation in or out of the liver. *NR1I2* (nuclear receptor subfamily 1, group I, member 2) and *NR1I3* (nuclear receptor subfamily 1, group I, member 3) encode proteins inducing *CYP450* enzymes and ABCB1.

This brief introduction of vitamin K related genes in human is updated and could be used as a reference in candidate identification (Figure 5.2).

5.4.2. *Calu* as a candidate gene associated with warfarin resistance

Our genotype-phenotype association test result for *Calu* gene in multiple wild populations designated *Calu* as a candidate gene affected by warfarin. The observed significant associations for the *Vkorc1*, *Calu* and its linked loci in both natural and the inferred virtual populations (Table 5.6) suggest the observation are not due to population structure or could be randomly obtained. Though lack of clear functional explanation for the observed polymorphism on this gene, previous study recognized the regulation role of *Calu* gene on the vitamin K-dependent g-carboxylation pathway (Markussen et al. 2007b). And the fact that *Calu* gene competitively bound the VKOR complex (encoded by the *Vkorc1* gene) with warfarin suggested an alternative way for developing resistance other than the direct variation on *Vkorc1* gene (Markussen et al. 2007a). The missing of association for *Vkorc1* gene and the detected association for *Calu* gene in the SP population gained our attention. This observation suggested alternative paths to warfarin resistance, at least in some rats. Moreover, the non-random association (LD) between *Vkorc1* and *Calu* reminds us of the potential role of *Calu* gene in warfarin-mediated pathways.

Chapter 6

Conclusions and Future directions

6.1. Conclusions

There is much hope to dissect the genetic architecture of adaptive traits in natural populations. However, as shown in this thesis, even in the case of what has been propagated as a classical system of a simple Mendelian adaptive trait the situation is more complex. Moreover, virtually all insights into the system date back to the 1970s and here it is for the first time that warfarin resistance in the rat is studied in the post genome era and with the knowledge of the warfarin resistance gene, *Vkorc1*, and the resistance causing mutation Y139C in hand.

My thesis work revealed that the main warfarin resistance gene *Vkorc1* set in the genetic background of rats from Germany likely is under strong (s~0.3) balancing selection; which confirms earlier inferences from field studies on phenotype frequencies but my work is novel in that it is the first population genetic analysis of the *Vkorc1* gene. Moreover, resistance in my study area likely has evolved from a new, as opposed to standing, genetic variant in *Vkorc1* (Y139C), which I inferred from forward time simulations that were most consistent with a starting frequency ~$1/N_e$ of the resistance allele around the time when resistance appeared in the area. This new mutation could be a *de novo* mutation or carried by single (or few) migrants (Chapter 2).

Beyond *Vkorc1*, we identified candidate genes related to warfarin resistance based on genomic variation data (Chapter 3) and microarray expression data (Chapter 4). A key innovation of my thesis is that I adopt a NetGWAS approach to analyze genomic data collected on wild populations. This approach has entered the area of complex disease gene mapping in the biomedical field, but has not been adopted by evolutionary

geneticists thus far. I show that while conventional association studies, or GWAS, hold much reduced potential to identify candidate genes related to warfarin resistance than does NetGWAS. A main reason for this is that adaptive loci in natural populations display strong haplotype structuring due to breeding structure and genetic hitchhiking; both leading to large numbers of false associations during single nucleotide polymorphism (SNP) genotyping and, interestingly and to my knowledge as not documented before, during gene expression analysis. The latter observation indicates that searches for adaptive trait loci in natural populations that have employed gene expression analyses potentially have dramatically overestimated the complexity of the genetic architecture of the adaptive trait under study, and/or potentially has misled interpretations of such results to the effect that the number of traits under selection was overestimated also. My combined analysis of SNP association with warfarin resistance and gene expression data showed that despite massive numbers of associated SNPs it is possible to distinguish between true association and false association due to linkage (and other factors) by adopting a NetGWAS approach. Furthermore, I complemented my analyses by conducting a *cis*-eQTL (expression quantitative trait loci) mapping approach, which substantiated my results.

Although we identified multiple candidate genes and selection signals on numerous chromosomes we were able to establish that analyzing these genes in a gene-gene network can reduce this apparent complexity of the genetics of warfarin resistance. Specifically, we were able to show that candidate genes can be connected to each other and shown to be centered on the vitamin K pathway; thereby forming what we could call a ‘warfarin resistance module’ (Figure 3.7A). We performed further analyses of

identified candidate genes in wild populations. The genes that have passed analyses described in the chapters 3-5 are summarized in Chapter 5. These are considered high-quality candidate genes that merit detailed further study. As a start, we collected genotype data and tested for the association of one such candidate gene, *Calu,* in 19 natural populations of rats where warfarin resistance is common. We observed that *Calu* is significantly associated with warfarin resistance in seven populations. It is interesting to note that any reference to the complexity of genetic architectures of adaptive traits appears to be manifest also in form of differences in the loci involved in different local populations. This would be analogous to genetic heterogeneity of human complex diseases. As such, I postulate a strong influence of genetic drift in shaping the complex genetics of adaptation to warfarin in our study area.

This network-guided genomic study demonstrated the importance of 'think globally', i.e. the consideration of gene-gene interactions in genomic studies. The NetGWAS and co-expression network analysis conducted here and supplemented by population genomics were shown as a powerful approach for the identification of candidate genes underlying warfarin resistance in the Norway rat. The evidence for selection at many of the identified genomic regions indicates that these regions indeed are part of a recent event.

6.2. Future directions

6.2.1. Is the *Vkorc1-Calu* interaction an example for the 'soft' selective sweep model underlying adaptation

Adaptations result from the rise in frequencies and fixation of beneficial alleles. This can occur by selection on a new mutant or by selection from standing variants (Pritchard and Di Rienzo 2010). As described by the 'selective sweep' model (c.f. Chapter 2), adjacent neutral alleles show similar evolutionary dynamics as the closely linked loci under selection (Kim and Stephan 2003). Traditionally, 'hard sweep' models refer to the situation that selection drives a new beneficial mutation to fixation rapidly, and genetic variants at linked sites are wiped out, thereby creating a large identical-by-descent (IBD) region (Pritchard, Pickrell, and Coop 2010). Detecting sweep signals thus became an effective means to detect potentially adaptive genes e.g. in human populations that evolved to deal with high altitude or benefitted from milk consumption (Enattah et al. 2008; Yi et al. 2010).

Recently, it has been advocated by the community that a polygenic, or 'soft sweep', model is might be more adequate to describe the process of genetic adaptation in natural populations than is the classical 'hard sweep' model (Pritchard and Di Rienzo 2010; Pritchard, Pickrell, and Coop 2010; Hernandez et al. 2011). This soft sweep model predicts small frequency shifts across many loci, and thus, under such 'polygenic adaptation' the fixation of beneficial alleles and the large IBD regions are not predicted. This would render genomic searches for such soft sweeps more difficult than searches for hard sweeps. In addition, under the polygenic model multiple loci with new or pre-

existing beneficial alleles respond to the selection directly by affecting the phenotype (Pritchard, Pickrell, and Coop 2010), or indirectly by interacting with the loci that interact with the trait. The population genetics of such interacting genes have not been explored to date, but likely follow soft sweep models also. Our study system of warfarin resistant rats and the now identified set of genes provide a unique opportunity for us to examine the soft sweep models in more detail.

Calu encodes an enzyme that binds to the VKOR complex. We have observed that *Calu* is associated with warfarin resistance and is in linkage disequilibrium with the resistance gene *Vkorc1* (Table 5.6, Table 5.7). As shown in Chapter 2, selection on *Vkorc1* created a sweep region spanning ~30 Mb around *Vkorc1*, which could be considered as a 'hard sweep' case except that the adaptive mutant is not fixed because of balancing selection. *Calu*, on the other hand, might be under a much weaker selection pressure and is part of a 'soft sweep' that is coupled with selection on *Vkorc1* even though the two genes are located on different chromosomes.

In addition, although the interferences between linked beneficial mutations have been noted and modeled in several studies (Kim and Stephan 2003; Illingworth and Mustonen 2011), genetic interactions between unlinked variants are not widely considered in population genetic theory. Now with the network perspective, we expect the gene-gene interactions between physically unlinked genes would potentially affect their evolutionary dynamics. Thus, both theoretical studies and empirical studies are needed that consider the population genetics of interacting genes underlying adaptive traits and the potential fitness cost associated with a soft sweep over numerous interacting

loci. We envision scenarios that merit exploration of the role of each gene in such groups of genes. For example, using the *Vkorc1-Calu* combination of genes one could distinguish between interactions of two genes that interact with warfarin ('potential drivers') as well as genes that merely hitchhike with these genes without direct contributions to the trait (hitchhikers).

6.2.2. Genes involved in arterial calcification – a fitness cost of resistance

The resistance mutation (Y139C) in *Vkorc1* confers tremendous advantage to rats when exposed to warfarin, but it also brings some fitness costs with it in form of reduced VKOR activity and, as was reported for a strain from the U.K., reduced growth rate (Smith, Townsend, and Smith 1991; Pelz et al. 2005; Rost et al. 2009). In a previous study of the wild-derived NW population rats displayed a cardiovascular phenotype (Kohn, Price, and Pelz 2008).

This emerging knowledge on the fitness cost of the adaptation to warfarin motivated us to search for genetic variants related to the phenotype of arterial calcification using our SNP array data. Provided that much evidence points to a higher cost of resistance in males than in females the role of sex-specific modifiers of the traits should be explored further.

A preliminary genomic association test has revealed nine SNPs that are significantly associated with the combined set of traits warfarin resistance, arterial calcification and gender. One of these nine SNPs is located in the 5' region of the *Fgfr2* gene (fibroblast growth factor receptor 2). This finding might indicate that this gene is

functionally related to our study variables, but the gene may be falsely associated with the traits because a selective sweep nearby at *Vkorc1* resulted in SNP associations over distances spanning 2 Mb separating the *Vkorc1* and *Fgfr2*. A previous GWAS on warfarin dose requirements in human patient cohorts observed that *FGFBP2* (fibroblast growth factor binding protein 2) contributed to the relatively high genotype-dose associations seen. Just as we observed here this was supported in addition to associations of *VKORC1* and *CYP2C9* (Cooper et al. 2008), and other loci discussed in my thesis. Thus, further investigations into *Fgfr2* association with resistance and a potential fitness cost are merited.

In addition, we found that the *Bglap* gene is significantly associated with both warfarin resistance and arterial calcification. With the prior knowledge of the dependence of this gene on vitamin dependent gamma-carboxylation and its association with mineralization of the bone matrix (Pan and Price 1985) we propose that the association detected in my thesis represents its potential role in vascular calcification seen in warfarin resistant Y139C mutant rats.

The Y139C mutation in *Vkorc1* is associated with arterial calcification phenotype with significance level P-value = 5.23×10^{-5}. We observed another SNP (S694261) with a significant P-value = 6.33×10^{-5} associated with arterial calcification. This SNP, located on chromosome 16 (position: 45349355 bp), is also associated with warfarin resistance before multiple correction (P-value = 0.021). There are 3 more SNPs in this region that exhibit such associations with both warfarin resistance and arterial calcification, however at lower levels of significance. Thus, an interesting question for future research refers to

the genes that map into this region. However, currently the resolutions of the SNP survey as well as the annotations of the region are too limited to resolve this question. As best as can be determined here there are two genes that map within 1mb of the S694261 SNP: *LOC100363103* and *LOC681082*.

These preliminary results revealed some interesting patterns associated with arterial calcification in rats, which might be induced by warfarin resistance and/or the exposure to warfarin. More comprehensive investigation including the network-guided analysis would help us understand the fitness costs of warfarin resistance in more detail, and could shed light on warfarin therapy in patients with chronic kidney disease given their high risk of vascular calcification.

6.2.3. Genetic architecture of resistance to second-generation anticoagulant rodenticides

Warfarin, a 4-hydroxycoumarin-based compound, is the first generation anticoagulant rodenticide. After rats resistant to warfarin had been found, other derivative of 4-hydroxycoumarin including coumatetralyl, bromadiolone, and difenacoum were introduced as rodent control compounds. We have obtained the phenotype data of rats resistant to some second-generation anticoagulant rodenticides (Kohn, Pelz, and Wayne 2003). For bromadiolone, we saw that *Vkorc1* gene and the flanking neutral alleles show significant associations ($P <= 0.055$) with the resistance phenotype in 4 natural populations. *Calu* and an upstream intergenic SNP are associated with the bromadiolone resistant phenotype in 3 natural populations. Interesting patterns of association were observed in the SP population, where all samples in this population are susceptible to

bromadiolone and resistance to warfarin was observed even though the Y139C mutation, or other *Vkorc1* mutations were not found. Thus, the observed resistance to warfarin in this population SP might be due to different mechanism. In addition, females were found to be more tolerant to second-generation anticoagulants than are males, and a previous study reported that sex is a factor affecting underlying resistance to bromadiolone (Kohn and Pelz 1999). Thus, we expect that the genetic architecture of resistance to bromadiolone is not limited to the known single resistance gene *Vkorc1*.

For the anticoagulant coumatetralyl we observed genotype-phenotype associations for *Vkorc1* and surrounding neutral SNPs in 4 natural populations, but none such association was observed for *Calu* and linked sites. Association of SNPs with difenacoum was not tested here because too few rats in the available sample are resistant to it. Resistance to this anticoagulant remains relatively sparse, e.g. as shown during a recent survey conducted in Belgium (Baert K 2012).

More detailed investigations of the genetics underlying resistance to second-generation anticogulant rodenticides are required. For example, previous studies have claimed that cytochrome P450 genes play an important role in bromadiolone resistance (Markussen et al. 2008a; Markussen et al. 2008b). Our study of warfarin resistance, however, did not support a statistical overrepresentation of cytochrome P450 genes. Thus, the role of cytochrome P450 genes in mediating bromadiolone resistance, as opposed to warfarin resistance, deserves further study.

Two additional areas of interest refer to the better quantification of the fitness costs of warfarin resistance and the specific genetic pathways that are pleiotropic; i.e. tat

can be shown to confer resistance but also can be shown to underlie detrimental traits. Finally, the fact that resistance has evolved from first- to second-generation anticogulants, called ‘resistance hierarchy’, poses a potential clue as to how adaptation progresses in nature (Pelz, Hänisch2, and Lauenstein 1995). Conceivably, in this system it is possible to infer in more detail how a set of mutations result in an adaptation to one environmental factor and to pre-adaptation to similar such factors.

6.2.4. The importance of adopting a gene-gene interaction network perspective

We have demonstrated the importance of adopting a gene-gene interaction network perspective when conducting GWAS and analyses of gene expression. As is shown in Figure 3.1, instead of identifying individual SNPs, under this framework SNPs are mapped to genes, and genes are mapped onto genetic networks prior to computing gene ranks. Besides the advantages of identification of individual genes that are associated with a trait this gene-gene interaction network perspective is helpful also in the context of the study of the genetic architecture underlying adaptive traits in that we were able to show that selection for warfarin resistance resulted in the association of clusters of highly connected genes with the trait. It might be feasible to adopt and modify approaches currently explored in the biomedical research field, where tools and algorithms aiming to test for associations between subnetworks with patient survival data have been developed (Li et al. 2012).

References

Adams, N., and R. Boice. 1983. A longitudinal study of dominance in an outdoor colony of domestic rats. Journal of Comparative Psychology 97:24-33.

Akula, N., A. Baranova, D. Seto et al. 2011. A network-based approach to prioritize results from genome-wide association studies. PLoS ONE 6:e24220-e24220.

Avery, P. J., A. Jorgensen, A. K. Hamberg, M. Wadelius, M. Pirmohamed, and F. Kamali. 2011. A proposal for an individualized pharmacogenetics-based warfarin initiation dose regimen for patients commencing anticoagulation therapy. Clin Pharmacol Ther 90:701-706.

Axelsson, E., A. Ratnakumar, M.-L. Arendt, K. Maqbool, M. T. Webster, M. Perloski, O. Liberg, J. M. Arnemo, Ö. Hedhammar, and K. Lindblad-Toh. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature.

Baert K, S. J. 2012. Distribution of anticoagulant resistance in the brown rat in Belgium. Belgian Journal of Zoology 142:39-48.

Baniasadi, S., S. Beizaee, B. Kazemi, N. Behzadnia, B. Shafaghi, M. Bandehpour, and F. Fahimi. 2011. Novel *VKORC1* mutations associated with warfarin sensitivity. Cardiovascular Therapeutics 29:e1-e5.

Barabasi, A.-L., N. Gulbahce, and J. Loscalzo. 2011. Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56-68.

Barabasi, A.-L., and Z. n. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5:101-113.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Barrett, R. D., and D. Schluter. 2008. Adaptation from standing genetic variation. Trends Ecol Evol 23:38-44.

Barrett, R. D. H., and H. E. Hoekstra. 2011. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12:767-780.

Beckmann-Knopp, S., S. Rietbrock, R. Weyhenmeyer, R. H. Bocker, K. T. Beckurts, W. Lang, M. Hunz, and U. Fuhr. 2000. Inhibitory effects of silibinin on cytochrome P-450 enzymes in human liver microsomes. Pharmacol Toxicol 86:250-256.

Benfey, P. N., and T. Mitchell-Olds. 2008. From genotype to phenotype: systems biology meets natural variation. Science 320:495-497.

Bollback, J. P., T. L. York, and R. Nielsen. 2008. Estimation of $2N_e s$ From Temporal Allele Frequency Data. Genetics 179:497-502.

Boyle, C. M. 1960. Case of apparent resistance of *Rattus norvegicus* Berkenhout to anticoagulant poisons. Nature 188:517-517.

Brachi, B., G. P. Morris, and J. O. Borevitz. 2011. Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12:232.

Brown, E. A. 2012. Genetic explorations of recent human metabolic adaptations: hypotheses and evidence. Biol Rev Camb Philos Soc 87:838-855.

Burmeister, M. 1999. Basic concepts in the study of diseases with complex genetics. Biol Psychiatry 45:522-532.

Campana, M. G., H. V. Hunt, H. Jones, and J. White. 2011. CorrSieve: software for summarizing and evaluating Structure output. Molecular Ecology Resources 11:349-352.

Camus-Kulandaivelu, L. t., J.-B. Veyrieras, B. Gouesnard, A. Charcosset, and D. Manicacci. 2006. Evaluating the reliability of outputs in case of relatedness between individuals.

Carvajal-Rodriguez, A. 2008. Simulation of genomes: a review. Current Genomics 9:155-159.

Cha, P.-C., T. Mushiroda, A. Takahashi, S. Saito, H. Shimomura, T. Suzuki, N. Kamatani, and Y. Nakamura. 2007. High-resolution SNP and haplotype maps of the human gamma-glutamyl carboxylase gene (*GGCX*) and association study between polymorphisms in *GGCX* and the warfarin maintenance dose requirement of the Japanese population. Journal of Human Genetics 52:856-864.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2.

Chen, Y.-L., B. Liu, Z.-N. Zhou, R.-Y. Hu, C. Fei, Z.-H. Xie, and X. Ding. 2009. Smad6 inhibits the transcriptional activity of *Tbx6* by mediating its degradation. The Journal of Biological Chemistry 284:23481-23490.

Clune, J., J.-B. Mouret, and H. Lipson. 2013. The evolutionary origins of modularity. Proceedings of the Royal Society B: Biological Sciences 280.

Cockram, J., J. White, F. J. Leigh, V. J. Lea, E. Chiapparino, D. A. Laurie, I. J. Mackay, W. Powell, and D. M. O'Sullivan. 2008. Association mapping of partitioning loci in barley. BMC Genetics 9:16-16.

Consortium, T. G. O. 2008. The Gene Ontology project in 2008. Nucleic Acids Research 36:D440-D444.

Consortium, T. I. H. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

Consortium, T. W. T. C. C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

Cooper, G. M., J. A. Johnson, T. Y. Langaee et al. 2008. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112:1022-1027.

Cordell, H. J. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10:392-404.

Cordell, H. J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463-2468.

Cornelis, M. C., K. L. Monda, K. Yu et al. 2011. Genome-wide meta-analysis identifies regions on 7p21 (*AHR*) and 15q24 (*CYP1A2*) as determinants of habitual caffeine consumption. PLoS Genet 7:e1002033-e1002033.

Coyne, J. 2009. Why evolution is true. Oxford University Press.

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, D. A. Nickerson, and M. Stephens. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700-706.

D.W.Marquardt. 1963. An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math 11:431-441.

Danziger, J. 2008. Vitamin K-dependent proteins, warfarin, and vascular calcification. Clin J Am Soc Nephrol 3:1504-1510.

Darwin, C. 1859. On the origin of species by means of natural selection, London.

Davis, N. A., J. E. Crowe, Jr., N. M. Pajewski, and B. A. McKinney. 2010. Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. Genes Immun 11:630-636.

Dixon, A. L., L. Liang, M. F. Moffatt et al. 2007. A genome-wide association study of global gene expression. Nature Genetics 39:1202-1207.

Earl, D., and B. vonHoldt. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4:359-361.

Ellegren, H., L. Smeds, R. Burri et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature.

Enattah, N. S., T. G. K. Jensen, M. Nielsen et al. 2008. Independent Introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. The American Journal of Human Genetics 82:57-72.

Endler, J. A. 1985. Natural selection in the wild. Princeton Univ. Press, Princeton.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14:2611-2620.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

Fu, H.-H., D. K. J. Lin, and H.-T. Tsai. 2006. Damping factor in Google page ranking: Research Articles. Appl. Stoch. Model. Bus. Ind. 22:431-444.

Gage, B. F., C. Eby, P. E. Milligan, G. A. Banet, J. R. Duncan, and H. L. McLeod. 2003. Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. Thrombosis and Haemostasis.

Gao, H., S. Williamson, and C. D. Bustamante. 2007. A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635-1651.

Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307-315.

Geisen, C., M. Watzka, K. Sittinger, M. Steffens, L. Daugela, E. Seifried, C. R. Müller, T. F. Wienker, and J. Oldenburg. 2005. VKORCI haplotypes and their impact on the inter-individual and inter-ethnical variability of oral anticoagulation. Thrombosis and Haemostasis.

Ghazalpour, A., S. Doss, B. Zhang et al. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2:e130.

Gilad, Y., S. A. Rifkin, and J. K. Pritchard. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics 24:408-415.

Gillespie, J. H. 1991. The causes of molecular evolution. Oxford Univ. Press, New York.

Goodstadt, L., and C. P. Ponting. 2004. Vitamin K epoxide reductase: homology, active site and catalytic mechanism. Trends in Biochemical Sciences 29:289-292.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8:195-202.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. The American Journal of Human Genetics 82:100-112.

Greaves, J. H., and P. Ayres. 1969. Linkages between genes for coat colour and resistance to warfarin in *Rattus norvegicus*. Nature 224:284-285.

Greaves, J. H., R. Redfern, P. B. Ayres, and J. E. Gill. 1977. Warfarin resistance: a balanced polymorphism in the Norway rat. Genetics Research 30:257-263.

Guan, Y., and M. Stephens. 2008. Practical issues in imputation-based association mapping. PLoS Genetics 4:e1000279-e1000279.

Hahn, M. W., M. V. Rockman, N. Soranzo, D. B. Goldstein, and G. A. Wray. 2004. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. Genetics 167:867-877.

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky, and A. D. Rienzo. 2010. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2459-2468.

Hans-Joachim, P., H. n. Detlef, and L. Gerhard. 1995. Resistance to anticoagulant rodenticides in Germany and future strategies to control *Rattus norvegicus*. Pesticide Science 43:61-67.

Harr, B., M. Kauer, and C. Schlötterer. 2002. Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences 99:12949-12954.

Hartl, D. L., and A. G. Clark. 2007. Principles of population genetics. Sinauer Associates, Inc.

Hedrick, P. W. 2012. What is the evidence for heterozygote advantage selection? Trends in Ecology & Evolution 27:698-704.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, P. Genomes, G. Sella, and M. Przeworski. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920-924.

Hoban, S., G. Bertorelle, and O. E. Gaggiotti. 2012. Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics 13:110-122.

Hohenlohe, P. A., S. Bassham, M. Currey, and W. A. Cresko. 2011. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philosophical Transactions of the Royal Society B: Biological Sciences 367:395-408.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Hurst, L. D. 2009. Genetics and the understanding of selection. Nature Reviews Genetics 10:83-93.

Ihaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299-314.

Illingworth, C. J. R., and V. Mustonen. 2011. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. Genetics 189:989-1000.

Imaoka, S., T. Hashizume, and Y. Funae. 2005. Localization of rat cytochrome P450 in various tissues and comparison of arachidonic acid metabolism by rat P450 with that by human P450 orthologs. Drug Metabolism and Pharmacokinetics 20:478-484.

Ivaskevicius, V., E. Jusciute, M. Steffens, C. Geisen, P. Hanfland, T. F. Wienker, E. Seifried, and J. Oldenburg. 2005. gammaAla82Gly represents a common fibrinogen gamma-chain variant in Caucasians. Blood Coagulation & Fibrinolysis: An International Journal in Haemostasis and Thrombosis 16:205-208.

Ivliev, A. E., V. A. Rudneva, and M. G. Sergeeva. 2010. Applicability of coexpression networks analysis to anticancer drug targets discovery. Molecular Biology 44:366-374.

IWPC. Warfarin Dosing. http://warfarindosing.org.

Jackson, W. B., and D. Kaukeinen. 1972. Resistance of wild Norway rats in north Carolina to warfarin rodenticide. Science 176:1343-1344.

Jacob, J., S. Endepols, H.-J. Pelz, E. Kampling, T. G. Cooper, C. H. Yeung, K. Redmann, and S. Schlatt. 2012. Vitamin K requirement and reproduction in bromadiolone-resistant Norway rats. Pest Management Science 68:378-385.

Jafari, P., and F. Azuaje. 2006. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Medical Informatics and Decision Making 6.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.

Jensen, L. J., M. Kuhn, M. Stark et al. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37:D412-416.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Research 14:528-538.

Johannsen, W. 1911. The genotype conception of heredity. The American Naturalist 45:129-159.

Kamali, F. 2006. Genetic influences on the response to warfarin. Current Opinion in Hematology 13:357-361.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27 - 30.

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513-1524.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777.

Kim, Y., and W. Stephan. 2003. Selective sweeps in the presence of interference among partially linked loci. Genetics 164:389-398.

Kim, Y., and T. Wiehe. 2009. Simulation of DNA sequence evolution under models of recent directional selection. Briefings in Bioinformatics 10:84-96.

Kohn, M. H., and H.-J. Pelz. 2000. A gene-anchored map position of the rat warfarin-resistance locus, Rw, and its orthologs in mice and humans. Blood 96:1996-1998.

Kohn, M. H., and H.-J. Pelz. 1999. Genomic assignment of the warfarin resistance locus, Rw, in the rat. Mammalian Genome 10:696-698.

Kohn, M. H., H. J. Pelz, and R. K. Wayne. 2003. Locus-specific genetic differentiation at Rw among warfarin-resistant rat (*Rattus norvegicus*) populations. Genetics 164:1055-1070.

Kohn, M. H., H. J. Pelz, and R. K. Wayne. 2000. Natural selection mapping of the warfarin-resistance gene. Proc Natl Acad Sci U S A 97:7911-7915.

Kohn, M. H., R. E. Price, and H.-J. Pelz. 2008. A cardiovascular phenotype in warfarin-resistant *Vkorc1* mutant rats. Artery research 2:138-147.

Kudaravalli, S., J. B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, and J. K. Pritchard. 2009. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol 26:649-658.

LaFramboise, T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Research 37:4181-4193.

Langfelder, P., and S. Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Larkin, M. A., G. Blackshields, N. P. Brown et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Lee, I., B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte. 2008. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat Genet 40:181-188.

Li, J., P. Roebuck, S. Grünewald, and H. Liang. 2012. SurvNet: a web server for identifying network-based biomarkers that most correlate with patient survival data. Nucleic Acids Research 40:W123-W126.

Li, T., C. Y. Chang, D. Y. Jin, P. J. Lin, A. Khvorova, and D. W. Stafford. 2004. Identification of the gene for vitamin K epoxide reductase. Nature 427:541-544.

Linnen, C. R., Y.-P. Poh, B. K. Peterson, R. D. H. Barrett, J. G. Larson, J. D. Jensen, and H. E. Hoekstra. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339:1312-1316.

Liu, J. Z., A. F. McRae, D. R. Nyholt et al. 2010. A versatile gene-based test for genome-wide association studies. Am J Hum Genet 87:139-145.

Liu, Y., H. Xu, S. Chen et al. 2011. Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases. PLoS Genet 7:e1001338.

Long, M., and L. Zhang. 2012. Why rodent pseudogenes refuse to retire. Genome Biology 13.

Lucia A., H., M. J, M. J, J. HA, H. PN, K. AK, and M. TA. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/.

Ludwig, A., M. Pruvost, M. Reissmann et al. 2009. Coat color variation at the beginning of horse domestication. Science 324:485-485.

Luna, A., and K. K. Nicodemus. 2007. snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. Bioinformatics 23:774-776.

Lund, M. 1964. Resistance to warfarin in the common rat. Nature 203:778-778.

Maertzdorf, J., D. Repsilber, S. K. Parida, K. Stanley, T. Roberts, G. Black, G. Walzl, and S. H. E. Kaufmann. 2011. Human gene expression profiles of susceptibility and resistance in tuberculosis. Genes Immun 12:15-22.

Manna, F., G. Martin, and T. Lenormand. 2011. Fitness landscapes: an alternative theory for the dominance of mutation. Genetics 189:923-937.

Markussen, M. D., A.-C. Heiberg, C. Alsbo, P. S. Nielsen, S. Kauppinen, and M. Kristensen. 2007a. Involvement of hepatic xenobiotic related genes in bromadiolone resistance in wild Norway rats, *Rattus norvegicus* (Berk.). Pesticide Biochemistry and Physiology 88:284-295.

Markussen, M. D., A. C. Heiberg, M. Fredholm, and M. Kristensen. 2007b. Characterization of bromadiolone resistance in a danish strain of Norway rats, *Rattus norvegicus*, by hepatic gene expression profiling of genes involved in vitamin K-dependent gamma-carboxylation. J Biochem Mol Toxicol 21:373-381.

Markussen, M. D. K., A.-C. Heiberg, M. Fredholm, and M. Kristensen. 2008a. Differential expression of cytochrome P450 genes between bromadiolone-resistant and anticoagulant-susceptible Norway rats: a possible role for pharmacokinetics in bromadiolone resistance. Pest Management Science 64:239-248.

Markussen, M. D. K., A.-C. Heiberg, M. Fredholm, and M. Kristensen. 2008b. Identification of cytochrome P450 differentiated expression related to developmental stages in bromadiolone resistance in rats (*Rattus norvegicus*). Pesticide Biochemistry and Physiology 91:147-152.

Markussen, M. D. K., A.-C. Heiberg, R. Nielsen, and H. Leirs. 2003. Vitamin K requirement in Danish anticoagulant-resistant Norway rats (*Rattus norvegicus*). Pest Management Science 59:913-920.

McClintick, J. N., and H. J. Edenberg. 2006. Effects of filtering by Present call on analysis of microarray experiments. BMC Bioinformatics 7:49.

McDonald, M. G., M. J. Rieder, M. Nakano, C. K. Hsia, and A. E. Rettie. 2009. *CYP4F2* Is a vitamin K1 oxidase: an explanation for altered warfarin dose in carriers of the V433M variant. Molecular Pharmacology 75:1337-1346.

Mezmouk, S., P. Dubreuil, M. Bosio, L. Décousset, A. Charcosset, S. Praud, and B. Mangin. 2011. Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik 122:1149-1160.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430:743-747.

Morrison, J. L., R. Breitling, D. J. Higham, and D. R. Gilbert. 2005. GeneRank: Using search engine technology for the analysis of microarray experiments. BMC Bioinformatics 6:233-233.

Neuditschko, M., M. S. Khatkar, and H. W. Raadsma. 2010. NetView: a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. PLoS ONE 7.

Nickerson, D. A., V. O. Tobe, and S. L. Taylor. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25:2745-2751.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15:1566-1575.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 27:29-34.

Oldenburg, J., C. G. Bevans, C. R. Muller, and M. Watzka. 2006. Vitamin K epoxide reductase complex subunit 1 (VKORC1): the key protein of the vitamin K cycle. Antioxid Redox Signal 8:347-353.

Orr, H. A. 2005. The genetic theory of adaptation: a brief history. Nature Reviews Genetics 6:119-127.

Orr, H. A., and S. Irving. 1997. The genetics of adaptation: the genetic basis of resistance to wasp parasitism in *Drosophila melanogaster*. Evolution 51:1877-1885.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: bringing order to the web.

Pan, L. C., and P. A. Price. 1985. The propeptide of rat bone gamma-carboxyglutamic acid protein shares homology with other vitamin K-dependent protein precursors. Proc Natl Acad Sci U S A 82:6109-6113.

Parker, H. G., B. M. VonHoldt, P. Quignon et al. 2009. An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325:995-998.

Partridge, G. G. 1979. Relative fitness of genotypes in a population of *Rattus norvegicus* polymorphic for warfarin resistance. Heredity 43:239-246.

Partridge, G. G. 1980. The vitamin K requirements of wild brown rats (*Rattus norvegicus*) resistant to warfarin. Comparative Biochemistry and Physiology Part A: Physiology 66:83-87.

Pautas, E., C. Moreau, I. Gouin-Thibault et al. 2009. Genetic factors (*VKORC1*, *CYP2C9*, *EPHX1*, and *CYP4F2*) are predictor variables for warfarin response in very elderly, frail inpatients. Clin Pharmacol Ther 87:57-64.

Pavani, A., S. M. Naushad, Y. Rupasree, T. R. Kumar, A. R. Malempati, R. K. Pinjala, R. C. Mishra, and V. K. Kutala. 2012. Optimization of warfarin dose by population-specific pharmacogenomic algorithm. The Pharmacogenomics Journal 12:306-311.

Pelz, H.-J., D. Hänisch2, and G. Lauenstein. 1995. Resistance to anticoagulant rodenticides in Germany and future strategies to control Rattus norvegicus. Pesticide Science 43:61-67.

Pelz, H.-J., S. Rost, M. Hünerberg et al. 2005. The genetic basis of resistance to anticoagulants in rodents. Genetics 170:1839-1847.

Peng, B., C. I. Amos, and M. Kimmel. 2007. Forward-time simulations of human populations with complex diseases. PLoS Genet 3:e47.

Peng, B., and M. Kimmel. 2005. simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21:3686-3687.

population, C. o. s. f. t. m. o. p. r. p., and B. o. agriculture. 1986. Pesticide Resistance. National Academy Press.

Presnell, S. R., and D. W. Stafford. 2002. The vitamin K-dependent carboxylase. Journal of Thrombosis and Haemostasis 87:937-946.

Pritchard, J. K., and A. Di Rienzo. 2010. Adaptation – not by sweeps alone. Nat Rev Genet 11:665-667.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology 20:R208-R215-R208-R215.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Purcell, S., B. Neale, K. Todd-Brown et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

Radwan, J., and W. Ç. Babik. 2012. The genomics of adaptation. Proceedings of the Royal Society B: Biological Sciences.

Raftogianis, R. B., T. C. Wood, and R. M. Weinshilboum. 1999. Human phenol sulfotransferases *SULT1A2* and *SULT1A1*: genetic polymorphisms, allozyme properties, and human liver genotype-phenotype correlations. Biochem Pharmacol 58:605-616.

Reja, V., A. Kwok, G. Stone, L. Yang, A. Missel, C. Menzel, and B. Bassam. 2010. ScreenClust: Advanced statistical software for supervised and unsupervised high resolution melting (HRM) analysis. Methods 50:S10-S14.

Rieder, M. J., A. P. Reiner, B. F. Gage, D. A. Nickerson, C. S. Eby, H. L. McLeod, D. K. Blough, K. E. Thummel, D. L. Veenstra, and A. E. Rettie. 2005. Effect of *VKORC1* haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352:2285-2293.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516-1517.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rosenberg, N. A. 2004. distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4:137-138.

Rost, S., A. Fregin, V. Ivaskevicius et al. 2004. Mutations in *VKORC1* cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427:537-541.

Rost, S., H.-J. Pelz, S. Menzel, A. D. MacNicoll, V. Leon, K.-J. Song, T. Jaekel, J. Oldenburg, and C. R. Muller. 2009. Novel mutations in the *VKORC1* gene of wild rats and mice - a response to 50 years of selection pressure by warfarin? BMC Genetics 10:4-4.

Runge-Morris, M., K. Rose, C. N. Falany, and T. A. Kocarek. 1998. Differential regulation of individual sulfotransferase isoforms by phenobarbital in male rat liver. Drug Metab Dispos 26:795-801.

Sabeti, e. a. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Saïdou, A.-A., C. Mariac, V. Luong, J.-L. Pham, G. Bezançon, and Y. Vigouroux. 2009. Association studies identify natural variation at PHYC linked to flowering time and morphological variation in Pearl Millet. Genetics 182:899-910.

Schadt, E. E., J. Lamb, X. Yang et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710-717.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78:629-644.

Schlenke, T. A., and D. J. Begun. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila Simulans*. Proceedings of the National Academy of Sciences of the United States of America 101:1626-1631.

Servin, B., and M. Stephens. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics 3:e114-e114.

Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485.

Smith, J. M., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetical Research 23:23-35.

Smith, P., M. Berdoy, R. H. Smith, and D. W. Macdonald. 1993. A new aspect of warfarin resistance in wild rats: benefits in the absence of poison. Functional Ecology 7:190-194.

Smith, P., M. G. Townsend, and R. H. Smith. 1991. A cost of resistance in the brown rat? Reduced growth rate in warfarin-resistant lines. Functional Ecology 5:441-447.

Stafford, D. W. 2005. The vitamin K cycle. Journal of Thrombosis and Haemostasis 3:1873-1878.

Stephan, W., Y. S. Song, and C. H. Langley. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172:2647-2663.

Stephens, M., and D. J. Balding. 2009. Bayesian statistical methods for genetic association studies. Nat Rev Genet 10:681-690.

Stephens, M., and N. Li. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213-2233.

Stephens, M., and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. American Journal of Human Genetics 76:449-462.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68:978-989.

Stranger, B. E., A. C. Nica, M. S. Forrest et al. 2007. Population genomics of human gene expression. Nature Genetics 39:1217-1224.

Stranger, B. E., E. A. Stahl, and T. Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187:367-383.

Suttie, J. W. 1993. Synthesis of vitamin K-dependent proteins. FASEB J. 7:445-452.

Tabrett, C. A., and M. W. Coughtrie. 2003. Phenol sulfotransferase 1A1 activity in human liver: kinetic properties, interindividual variation and re-evaluation of the suitability of 4-nitrophenol as a probe substrate. Biochem Pharmacol 66:2089-2097.

Takeuchi, F., R. McGinnis, S. Bourgeois et al. 2009. A genome-wide association study confirms *VKORC1*, *CYP2C9*, and *CYP4F2* as principal genetic determinants of warfarin dose. PLoS Genet 5:e1000433.

Team, R. D. C. 2012. R: A language and environment for statistical computing.

Thijssen, H. H. W. 1995. Warfarin-based rodenticides : mode of action and mechanism of resistance. Pesticide Science:73-78.

Thomas, J. H. 2007. Rapid birth-death evolution specific to xenobiotic Cytochrome P450 genes in vertebrates. PLoS Genet 3.

Tishkoff, S. A., F. A. Reed, A. Ranciaro et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31-40.

UI-Health. 2012. A team approach to warfarin pharmacogenetics. Pharmacy Practice News 39.

van Noort, V., B. Snel, and M. A. Huynen. 2003. Predicting gene function by conserved co-expression. Trends in Genetics 19:238-242.

Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. Springer, New York.

Veyrieras, J.-B., S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214-e1000214.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol 4:e72.

Wadelius, M., L. Y. Chen, K. Downes et al. 2005. Common *VKORC1* and *GGCX* polymorphisms associated with warfarin dose. Pharmacogenomics J 5:262-270.

Wadelius, M., L. Y. Chen, N. Eriksson, S. Bumpstead, J. Ghori, C. Wadelius, D. Bentley, R. McGinnis, and P. Deloukas. 2007. Association of warfarin dose with genes involved in its action and metabolism. Human Genetics 121:23-34.

Wadelius, M., L. Y. Chen, J. D. Lindh, N. Eriksson, M. J. R. Ghori, S. Bumpstead, L. Holm, R. McGinnis, A. Rane, and P. Deloukas. 2009. The largest prospective warfarin-treated cohort supports genetic forecasting. Blood 113:784-792.

Wajih, N., D. C. Sane, S. M. Hutson, and R. Wallin. 2004. The inhibitory effect of calumenin on the vitamin K-dependent gamma-carboxylation system. Characterization of the system in normal and warfarin-resistant rats. J Biol Chem 279:25276-25283.

Wallace, M. E., and F. M. MacSwiney. 1979. An inherited mild middle-aged adiposity in wild mice. The Journal of Hygiene 82:309-317.

Wallin, R., S. M. Hutson, D. Cain, A. Sweatt, and D. C. Sane. 2001. A molecular mechanism for genetic warfarin resistance in the rat. FASEB J. 15:2542-2544.

Wang, B., Y.-B. Zhang, F. Zhang et al. 2011. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS ONE 6.

Wang, Y., U. Wimmer, P. Lichtlen et al. 2004. Metal-responsive transcription factor-1 (MTF-1) is essential for embryonic liver development and heavy metal detoxification in the adult liver. FASEB J 18:1071-1079.

Wang, Y., L. P. Zhao, and S. Dudoit. 2006. A fine-scale linkage-disequilibrium measure based on length of haplotype sharing. American Journal of Human Genetics 78:615-628.

White, H. E., V. J. Hall, and N. C. P. Cross. 2007. Methylation-sensitive high-resolution melting-curve analysis of the *SNRPN* Gene as a diagnostic screen for Prader-Willi and Angelman syndromes. Clin Chem 53:1960-1962.

Winter, C., G. Kristiansen, S. Kersting et al. 2012. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput Biol 8:e1002511.

Wittkopp, P. J. 2007. Variable gene expression in eukaryotes: a network perspective. Journal of Experimental Biology 210:1567-1575.

Wittwer, C. T., G. H. Reed, C. N. Gundry, J. G. Vandersteen, and R. J. Pryor. 2003. High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem 49:853-860.

Wu, Z., R. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. 2004. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99:909.

Xiong, M., and S. W. Guo. 1997. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. American Journal of Human Genetics 60:1513-1531.

Yanagita, M. 2004. *Gas6*, warfarin, and kidney diseases. Journal of Clinical and Experimental Nephrology 8:304-309.

Yi, X., Y. Liang, E. Huerta-Sanchez et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75-78.

Zebboudj, A. F., V. Shin, and K. Bostrom. 2003. Matrix GLA protein and BMP-2 regulate osteoinduction in calcifying vascular cells. Journal of Cellular Biochemistry 90:756-765.

Appendix

Appendix 1 – Population and sample information.

Farm No.	Farm Name	Sample size	R&S(N)	R(N)	R(%)	Rc/Rc+ (N)	Rb/Rb+(N)	Rd/Rd+ (N)	Abbre vation
-	**NW**	**29**	**29**	**20**	**68.9%**	**-**	**-**	**-**	**NW**
4f	**LUDWIGSHAFEN**	**13**	**13**	**0**	**0%**	**9**	**0**	**0**	**LH**
24	**WANNING-UPGANG**	73	55	51	92.7%	56	51	43	WU
11	**KORTENBUSCH**	70	65	54	83.1%	30	51	39	KB
21	**SCHULTE-SPECHTEL**	23	21	7	33.3%	3	14	0	SP
23&16	**THISSEN & NARMANN**	56	55	4	7.3%	17	7	2	TH
19	**PIEPER**	56	49	32	65.3%	28	33	23	
17	**NIEHOFF**	53	51	44	86.3%	44	42	37	
4	**BEUCK**	52	46	43	93.5%	21	37	31	
10	**KLEVERTH**	42	38	36	94.7%	23	32	28	
13	**LEPPMANN**	42	36	23	63.9%	27	25	19	
28	**WESTLINNING**	37	35	30	85.7%	20	29	21	
18	**NOTTELMANN**	29	23	21	91.3%	8	19	15	
20	**SCHULTE-EVERDING**	27	26	16	61.5%	14	15	14	
12	**KOSTERS-BORKMAN**	25	21	20	95.2%	8	17	17	
6	**GUNNIGMANN**	24	24	23	95.8%	11	20	16	
5	**GROSSE-HOLZ**	20	19	14	73.7%	0	15	13	
25	**VERNAUER**	20	19	17	89.5%	13	15	12	
14	**LOOZ**	19	16	7	43.8%	3	12	4	
15	LUTTGE-HOLZ	10	6	1	16.7%	0	3	0	
30	DUISENBURG/ EMSLA	9	9	7	77.8%	9	7	5	
26	WAULIGMANN	6	6	6	100%	2	6	6	
29	BRAMHAR	6	6	4	66.7%	5	4	2	
3	BAYER-EMMERICH	5	4	1	25.0%	0	3	1	
32	LAHDEN	4	4	1	25.0%	2	2	0	
2	ASBECK/ SCHOPPINGEN	3	3	3	100%	0	3	1	
3f	MAGDEBURG/ZOO	3	3	1	33.3%	0	0	0	

1	ALTENAU	2	2	2	100%	2	2	2
8	HOLTKAMP	2	2	2	100%	2	2	2
27	WEHLING	2	2	2	100%	2	2	2
33	LUTCKE/ GIEVENBECK	2	2	1	50.0%	1	1	1
2f	KREIS SOEST	2	2	2	100%	2	2	2
7	HAMM-UENTROP	1	0	0	NA	1	0	0
9	KLEMANN	1	1	0	0%	1	0	0
22	SENDENHORST	1	1	1	100%	0	1	1
31	HAARANNEN/ FURST	1	1	1	100%	0	0	0
5f	JANSSEN/ SOLTBORG	1	1	0	0%	1	0	0

Populations (N > 10) were highlighted bold. NW: wild-derived; LH: non-resistant population.
R&S(N) The number of rat samples with the warfarin resistant/susceptible phenotype.
R(N): The number of rat samples with the warfarin resistant phenotype.
R(%) The warfarin resistant level: R(N) divided by R&S(N).
Coumatetralyl, bromadiolone and difenacoum are anticoagulant rodenticides.
Rc/Rc+(N) The number of rats with the coumatetralyl resistant/susceptible phenotype.
Rb/Rb+(N) The number of rats with the bromadiolone resistant/susceptible phenotype.
Rd/Rd+(N) The number of rat samples with the difenacoum resistant/susceptible phenotype.

Appendix 2 – Designed primers for PCR and High resolution melting genotyping.

I. Polymerase chain reaction (PCR)

Primer for SNP	Chr:Pos	PCR Forward	PCR Reservse
rs105465027	5:47202315	**CACGACGTTGTAAAACGAC**T TACAAACAAGTCGGCCAAG	**GGATAACAATTTCACACAGG**CA CGAGAGAGTACAGAGCCATTT
rs105163021	4:138458155	**CACGACGTTGTAAAACGAC**A CACACACCTCTACCACACCA	**GGATAACAATTTCACACAGG**CA TGACCTTGAGGAGGATGC
rs105210758	2:7301927	**CACGACGTTGTAAAACGAC**A GAAGCAAAACCCAATGAACA	**GGATAACAATTTCACACAGG**TG GTGTTGCCTGTGTCCTAA
rs197161785	1:11546996	**CACGACGTTGTAAAACGAC**C CTCCTCTTCACGCAACTTT	**GGATAACAATTTCACACAGG**TT TGGTGCTTTCCCCAGTAA
Primers for SNP discovery		**PCR Forward**	**PCR Reservse**
Rn_*Calu*_5' intergenic region		**CACGACGTTGTAAAACGAC**A AATGTTCCTTTGGCCTCCT	**GGATAACAATTTCACACAGG**CT TGGATCGGGAAAATGAAA
Rn_*Calu*_5' exon1		**CACGACGTTGTAAAACGAC**A GAAGTGTGAAAGACCAACGTG	**GGATAACAATTTCACACAGG**AG AACAAGGACGTGCGAAAC
Rn_*Calu*_5' intron1		**CACGACGTTGTAAAACGAC**C CTAGAGAATCCCCCACACA	**GGATAACAATTTCACACAGG**AA GGAGGAAGGAGGTGGAAA
Rn_*Calu*_5' exon2		**CACGACGTTGTAAAACGAC**G ACCTGTGTTATGGGGCAAG	**GGATAACAATTTCACACAGG**AT CCCCAACCCCTTCCTACT
Rn_*Calu*_5' exon3		**CACGACGTTGTAAAACGAC**G CTTTGCTGGTTTGCCTTAC	**GGATAACAATTTCACACAGG**AG GACTTCAGCTCCCCCTTC
Rn_*Calu*_5' exon4&5		**CACGACGTTGTAAAACGAC**G GTTAGAGATGAGCGGAGGTT	**GGATAACAATTTCACACAGG**AA GGTGTGTGCATGTATGAGTG
Rn_*Calu*_5' exon6		**CACGACGTTGTAAAACGAC**T CCCTGTCTGTGGAGATGTG	**GGATAACAATTTCACACAGG**GG CCTCTGCATGGTCATAGT
Rn_*Calu*_5' exon7		**CACGACGTTGTAAAACGAC**G CTCACCAAGGAGGAGATTG	**GGATAACAATTTCACACAGG**GG TGCCCAAATCGTTTATCT

II. High resolution melting (HRM) genotyping

Primer for	Chr:Pos	HRM Forward	HRM Reservse
Rn_*Vkorc1*_Y139C	1:187177049	TGTCTGTCGCTGGTTCTCTGT	AGGGCTTTTTGACCTTGTGTT
Rn_*Calu*_56228735	4:56228735	CTGTCGGCCTTGTTCTCTATTC	CGGGCTACAACCCTTTCC
Rn_*Calu*_56038243	4:56038243	TCAGCCCAGAGTAAGGATGC	GGGAGGGTAAGAGTCCTGGT
rs105465027	5:47202315	CCGGGAAAGGGAATAACAAT	AGGAGCCACTTGAACAAACG
rs105163021	4:138458155	AGCCACATGCTTTCTCAAAT	GCCCCTTGAGAGTACAGCAA
rs105210758	2:7301927	ATTGGCAAAATGAACCATTACAC	TCCTAAAGTTCCTATTTTCAGTCCA
rs197161785	1:11546996	CTCCTCTTCACGCAACTTTCTC	AGGCCCCTGAAGTGAATTTTAT
S422152	6:136715130	ATGTGCTACCCGAGTCTTCCT	GCTGTGTGGCCTTAAATAGCA
S693558	5:30306799	CAAGAGCTGGTTTGAGACAACA	TGAAAACAATCATGGGCAAA
S694902	4:63592604	AGAATGTCCCTGTGTGATTG	GCTTCACCAAACATCCCAGT
S696656	10:19930191	GCTGTGAGGGGAAACAAGAA	GCACCAATCTCCATCCACTT

Chr:Pos indicates the position of the SNP site on chromosome based on NCBI build 4.

Appendix 3 – Genotype-phenotype association tests of *Vkorc1* and surrounding region for NW samples (Figure 2.2).

SNP	Pos	Assoc_P -value	$\log_{10}$(BF1)	$\log_{10}$(BF2)	$\log_{10}$(BF3)	LH_linkfreq	NW_linkfreq(original)	NW_linkfreq (recovered)
S422912	172020573	4.03E-1	-0.076	-0.043	-0.055	0.792	0.725	0.745
S424296	172021329	8.23E-2	0.361	0.411	0.671	0.708	0.875	0.819
S425392	172199898	4.03E-1	-0.072	-0.045	0.013	0.708	0.725	0.745
S425775	172504303	1.21E-1	0.42	0.383	0.763	0.750	0.900	0.855
S426319	172577432	1.21E-1	0.373	0.356	0.756	0.750	0.900	0.855
S422189	172998518	4.29E-2	0.659	0.702	1.215	0.667	0.850	0.782
S693423	173147475	1.03E-1	-0.039	0.028	0.081	0.958	0.800	0.854
S425346	173398184	2.03E-2	0.8	1.037	1.609	0.500	0.800	0.710
S693021	173583296	3.11E-1	-0.067	-0.077	-0.196	0.909	0.800	0.819
S692583	173785040	7.57E-1	-0.038	-0.08	-0.186	0.667	0.750	0.747
S426162	174129878	3.11E-1	-0.074	-0.084	-0.173	0.583	0.800	0.819
S426183	174184945	4.86E-1	-0.068	-0.091	-0.266	0.667	0.816	0.825
S423031	174238208	6.72E-1	-0.038	-0.064	-0.235	0.583	0.763	0.766
S422569	174412169	2.68E-1	-0.069	-0.067	-0.119	0.708	0.825	0.855
S422184	174592784	2.31E-1	0.215	0.203	0.341	0.500	0.925	0.891
S424830	174638055	6.57E-1	0.018	0.048	-0.022	0.417	0.900	0.884
S423531	174804498	3.19E-1	-0.061	-0.026	-0.073	0.900	0.763	0.798
S694817	174957949	3.71E-1	0.101	0.138	0.164	0.900	0.969	0.991
S422107	175233123	1.15E-1	0.167	0.115	0.345	0.542	0.600	0.531
S421541	175536967	2.38E-1	0.17	0.19	0.344	0.958	0.950	0.927
S424580	175619519	6.78E-3	0.886	1.245	1.892	0.250	0.775	0.673
S693438	175677908	6.78E-3	1.029	1.277	1.881	0.250	0.775	0.673
S421567	175944948	6.13E-1	-0.077	-0.032	-0.052	0.625	0.725	0.742
S426628	176403842	5.13E-1	0.134	0.138	0.236	0.900	0.975	0.964
S693441	176588782	1.45E-2	0.198	0.361	0.292	0.833	0.650	0.748
S692235	176761645	1.32E-1	-0.057	0.118	0.41	0.773	0.789	0.842

S421516	177843913	2.31E-1	0.191	0.211	0.34	0.708	0.925	0.891
S696928	179327802	1	0.062	0.071	0.125	0.708	1.000	1.000
S426284	179862388	3.23E-3	0.412	0.382	0.557	0.727	0.700	0.820
S424953	180822160	1.02E-2	-0.02	0.096	0.264	0.900	0.900	0.964
S421720	181012647	7.34E-3	0.444	0.525	0.892	0.125	0.725	0.630
S423104	181287248	4.38E-1	0.009	-0.077	-0.331	0.100	0.450	0.484
S422575	181494720	3.17E-1	0.109	0.137	0.258	0.750	0.975	0.964
S420883	181818332	9.65E-2	0.153	0.294	0.422	0.900	0.725	0.658
S692233	182078913	9.65E-2	0.185	0.284	0.532	0.875	0.725	0.658
S424697	182145797	3.29E-2	0.386	0.569	0.864	0.167	0.325	0.415
S693902	183065176	2.61E-1	0.025	0.106	0.162	0.958	0.800	0.766
S421196	184360555	2.00E-1	-0.002	0.03	-0.012	0.792	0.700	0.760
S425055	184935644	1.50E-1	-0.03	0.01	0.067	0.750	0.842	0.893
S423171	184995554	5.13E-1	0.082	0.14	0.231	0.833	0.975	0.964
S694795	185341704	1.72E-1	0.258	0.286	0.42	0.750	0.925	0.891
S421721	185633136	1.93E-2	0.013	0.252	0.441	0.667	0.850	0.921
Sult1a1_ 185832744	185832744	1.44E-2	0.38	0.546	0.941	0.591	0.879	0.928
S421034	185853325	5.13E-1	0.17	0.141	0.234	0.909	0.975	0.964
S425356	186055581	1.93E-2	-0.001	0.172	0.444	0.833	0.850	0.921
S693022	186370915	5.03E-3	0.134	0.291	0.776	0.708	0.875	0.957
Spn_ 186389319	186389319	1.48E-2	0.301	0.452	0.837	0.682	0.914	0.957
Spn_ 186389641	186389641	4.60E-1	-0.008	-0.015	-0.053	0.864	0.983	0.977
S426028	186535990	1	-0.03	-0.011	-0.083	0.958	1.000	1.000
S693918	186703048	1.93E-2	0.055	0.164	0.427	0.625	0.850	0.921
Vkorc1_ 187177049	187177049	1.18E-6	4.158	6.222	8.235	0.000	0.483	0.636
Bckdk_ 187189027	187189027	4.60E-1	-0.008	-0.015	-0.053	0.900	0.983	0.977
Tgfb1i1_ 187506367	187506367	5.88E-4	0.835	1.318	2.12	0.818	0.914	0.973
Tgfb1i1_ 187506630	187506630	5.88E-4	0.835	1.318	2.12	0.818	0.914	0.973
Tgfb1i1_ 187506650	187506650	5.88E-4	0.835	1.318	2.12	0.818	0.914	0.973
S421312	187528619	1.34E-1	0.499	0.644	1.018	0.900	0.917	0.873
S691557	187787779	1.81E-1	0.007	0.101	0.333	0.636	0.750	0.804
Bag3_ 187802465	187802465	1.80E-3	0.484	0.777	1.393	0.818	0.931	0.980
Bag3_ 187802630	187802630	1.80E-3	0.484	0.777	1.393	0.818	0.931	0.980
Bag3_ 187802733	187802733	1.80E-3	0.484	0.777	1.393	0.818	0.931	0.980
Bag3_ 187802803	187802803	1.80E-3	0.484	0.777	1.393	0.818	0.931	0.980
Bag3_ 187802804	187802804	2.16E-1	0.091	0.111	0.144	0.636	0.724	0.760
Bag3_ 187802864	187802864	5.88E-4	0.835	1.318	2.12	0.818	0.914	0.973
S424424	187947190	1	0.067	0.082	0.142	0.875	1.000	1.000
Ppapdc1a_ 188588783	188588783	1.48E-2	0.301	0.452	0.837	0.909	0.914	0.955
Ppapdc1a_ 188588940	188588940	1.48E-2	0.301	0.452	0.837	0.900	0.914	0.955
Ppapdc1a_ 188588959	188588959	1.48E-2	0.301	0.452	0.837	0.900	0.914	0.955

Ppapdc1a_188589203	188589203	6.84E-4	1.251	1.129	1.539	0.091	0.603	0.709
Ppapdc1a_188589271	188589271	1.44E-2	0.457	0.375	0.506	0.227	0.517	0.596
Ppapdc1a_188589288	188589288	1	-0.005	-0.011	-0.028	0.955	1.000	1.000
S693024	188627443	1	-0.054	-0.064	-0.182	0.650	1.000	1.000
S695593	189396078	7.10E-4	1.758	1.782	2.393	0.083	0.375	0.516
S692253	190281302	1	0.064	0.083	0.15	0.958	1.000	1.000
S424949	191240400	7.71E-3	0.904	0.868	1.251	0.125	0.300	0.407
S692250	191697937	7.14E-3	0.256	0.126	0.32	0.150	0.600	0.702
S693026	191854465	2.47E-2	0.001	0.205	0.509	0.208	0.579	0.647
S422227	192027314	3.17E-1	-0.026	-0.038	-0.124	0.875	0.975	0.993
S692209	192027314	3.17E-1	-0.018	-0.05	-0.111	0.875	0.975	0.993
S694323	192700265	1	0.048	0.073	0.14	0.792	1.000	1.000
S425868	193218073	1.02E-2	-0.037	0.11	0.301	0.900	0.900	0.964
S422139	193248816	2.20E-2	0.206	0.231	0.422	0.364	0.650	0.743
S422001	193421868	3.17E-1	0.109	0.126	0.196	0.900	0.975	0.964
S693895	194505012	6.49E-1	0.008	-0.076	-0.013	0.583	0.632	0.605
S692595	194673191	8.41E-1	-0.074	-0.122	-0.383	0.417	0.625	0.622
S423289	195132894	1	0.041	0.076	0.155	0.958	1.000	1.000
S693428	195273115	1.35E-3	0.328	0.202	0.34	0.773	0.563	0.690
S424902	196018672	3.42E-1	0.177	0.213	0.355	0.875	0.950	0.927
S424023	196204823	5.13E-1	0.092	0.119	0.209	0.900	0.975	0.964
S694816	196572137	7.20E-3	0.392	0.326	0.519	0.500	0.694	0.594
S422355	197212851	8.66E-2	0.346	0.351	0.722	0.875	0.900	0.855
S422488	197533634	1.82E-2	0.197	0.232	0.546	0.455	0.639	0.734
S422095	197862317	7.30E-1	-0.059	-0.05	-0.099	0.708	0.625	0.628
S425251	197922583	1.48E-1	0.227	0.264	0.417	0.773	0.825	0.774
S425620	198161874	1.48E-1	0.274	0.261	0.346	0.833	0.825	0.774
S693884	199166234	7.84E-3	0.892	1.064	1.555	0.792	0.325	0.443
S423722	199384293	3.42E-1	0.22	0.189	0.272	0.375	0.950	0.927
S423093	199652599	1.94E-1	0.105	0.113	0.174	0.350	0.550	0.487
S422067	199983879	2.21E-3	1.016	0.903	1.423	0.900	0.425	0.559
S693421	200884723	1	0.069	0.06	0.095	0.875	0.725	0.710
S423813	201127302	6.12E-1	-0.02	-0.058	-0.181	0.958	1.000	1.000

SNP	The SNP ID. For SNPs from 10K Affymetrix array, we used the Assay ID; for SNPs genotyped by ourselves, we used the gene name and the physical position on chromosome 1
Pos	The physical position of each SNP on chromosome 1
Assoc_P-value	The unadjusted P values calculated using Cochran-Mantel-Haenszel (CMH) test controlling for the potential confounding factor sex
$\log_{10}$(BF1)	The $\log_{10}$ values of the Bayes Factor 1 using codominant genetic model
$\log_{10}$(BF2)	The $\log_{10}$ values of the Bayes Factor 2 using dominant genetic model
$\log_{10}$(BF3)	The $\log_{10}$ values of the Bayes Factor 3 using overdominant genetic model
LH_linkfreq	The frequency of the allele initially linked to *Vkorc1*'s mutation in LH population
NW_linkfreq(original)	The frequency of the allele initially linked to *Vkorc1*'s mutation for NW samples
NW_linkfreq(recovered)	The frequency of the allele initially linked to *Vkorc1*'s mutation in the recovered NW population

Appendix 4 – Candidate genes identified from SNP array I (Chapter 3).

Gene	Candidate SNPs	Chr	Gene Start	Gene End	Region	Support Score	GeneScore ($Log_{10}BF1$)	GeneScore ($Log_{10}BF2$)	NodeDegree (Net1)	NodeDegree (Net2)	LD blocks	iHS (XP-EHH)
Npas1	S693005	1	76719188	76738565	1	5	1.29	1.69	NA	18		
Ap2s1	S693005	1	77069125	77081190	1	5	1.29	1.69	12	25		
Parva	S424447: S426027	1	170232441	170387843	2	11	1.29	1.70	13	26	LD1	2.0
Tead1	S424447: S426027	1	170557393	170695818	2	5	1.29	1.70	17	28	LD1	2.0
Xylt1	S424580: S693438	1	175687360	175802134	3	6	1.13	1.36	7	12	LD1	2.0
Vkorc1	Vkorc1_ 187177049	1	187175098	187179262	4	12	4.16	6.22	9	12	LD1	
Fgfr2	S695593	1	189482977	189589279	4	6	1.64	1.91	NA	192	LD1	
Bche	S424818	2	164329613	164427994	5	12	1.30	1.74	50	85		
Tdo2	NA	2	173594018	173611833	6	7	0.07	0.12	182	323		
Bglap	S423241	2	180482313	180483290	7	7	0.05	0.00	152	274		2.6
Fdps	S423241	2	181168903	181177792	7	1	0.61	0.69	16	29		2.6
Tuft1	NA	2	189595502	189641602	8	1	0.61	0.89	7	14		
Wdr3	S425777: S426471	2	195093948	195116591	9	6	1.15	1.75	17	30		3.0
F3	S424133: S425042	2	218371050	218382644	10	3	1.03	1.77	NA	100		
Abcd3	S424133: S425042	2	218396081	218432176	10	3	1.03	1.77	NA	43		
Bcar3	S424133: S425042	2	219075206	219188344	10	5	1.04	1.37	14	21		
Unc5c	S425671	2	239568542	239721231	11	7	1.30	1.25	7	10		2.1
Bmpr1b	S425671	2	239727570	239777074	11	10	1.30	1.25	21	48		2.1
Ppig	NA	3	51866989	51895890	12	8	0.08	0.19	36	55	LD2	2.5
Olr672	S424398: S424527	3	73233335	73234249	13	5	1.12	1.31	1	1		
Olr673	S424398: S424527	3	73309447	73310382	13	5	1.12	1.31	1	1		

Olr674	S424398: S424527	3	73318353	73319288	13	5	1.12	1.31	1	1	
F2	S424527	3	76005320	76018603	13	8	0.18	0.21	58	105	
Shf	S422383	3	109129839	109149342	14	5	1.11	1.44	9	18	
Ascc3l1	S425673	3	114708625	114738537	15	5	1.03	1.31	48	86	
Ciao1	S425673	3	114740677	114746203	15	5	1.03	1.31	17	29	
Tmem127	S425673	3	114746454	114759246	15	4	1.03	1.31	5	8	
Stard7	S425673	3	114768201	114795440	15	4	1.03	1.31	2	4	
Rin2	S421464	3	134307538	134503915	16	12	2.06	2.44	6	9	
Nat5	S421464	3	134522185	134537185	16	12	2.06	2.44	3	8	
Sfrs6	S423582	3	153795483	153799024	17	6	1.72	2.42	NA	48	LD3
L3mbtl	S423582	3	153811334	153836990	17	6	1.72	2.42	NA	1	LD3
Sgk2	S423582	3	153849825	153890111	17	6	1.72	2.42	NA	11	LD3
Ift52	S423582	3	153893164	153917594	17	12	1.72	2.42	9	14	LD3
Mybl2	S423582	3	153925908	153954323	17	12	1.72	2.42	21	45	LD3
Prex1	S424732: S424154: S423483	3	157693897	157863943	17	12	2.24	2.97	10	21	LD3
Arfgef2	S424732: S424154: S423483	3	157967942	158051358	17	5	1.06	1.28	10	17	LD3
Cse1l	S424732: S424154: S423483	3	158061678	158100065	17	11	1.06	1.28	27	51	LD3
Stau1	S424732: S424154: S423483	3	158101209	158147105	17	5	1.06	1.28	21	40	LD3
Calu	NA	4	56228617	56256117	18	8	0.47	0.56	35	51	
RGD1565690	S424027	4	88376454	88378742	19	6	1.71	2.07	4	5	
Ggcx	NA	4	105719323	105735190	20	8	0.23	0.17	15	24	LD4
Mgp	NA	4	173910585	173913947	21	8	-0.07	-0.03	58	118	
Clta	S422841	5	60484404	60502432	22	5	1.13	1.30	61	103	LD5
Gne	S422841	5	60506174	60546285	22	6	1.13	1.30	20	37	LD5

Tex10	S694477	5	65118575	65165336	22	11	1.50	1.75	4	4	LD5	
RGD1308165	S694477	5	65226927	65249757	22	5	1.50	1.75	NA	4	LD5	
Tmeff1	S694477	5	65271278	65357295	22	6	1.50	1.75	NA	47	LD5	
Murc	S694477	5	65357777	65365621	22	5	1.50	1.75	NA	3	LD5	
Akap2	S422312	5	75814965	75855490	23	2	0.79	1.38	4	14		
Orm1	NA	5	80325042	80328194	24	4	0.02	0.00	NA	112		
Rock2	S426589: S421572	6	40581249	40672854	25	7	1.27	1.19	49	84		2.4
Rnf144a	S425096	6	43957013	44073940	25	2	1.03	1.07	2	3		2.4
Rsad2	S425096	6	44101919	44113911	25	3	1.03	1.07	14	28		2.4
Cmpk2	S425096	6	44128537	44139346	25	3	1.03	1.07	23	32		2.4
Tgfb3	S423641	6	110173444	110195215	26	5	1.00	1.30	70	122		
RGD1565705	S692740	6	143566620	143657319	27	2	1.19	1.25	NA	2		
Ptprn2	S692740	6	143787884	144549042	27	6	1.19	1.25	28	45		
Cyp4f1	NA	7	13589664	13600856	28	8	0.25	0.30	26	39		
RGD1566296	S424184	8	64149487	64350969	29	4	1.10	1.19	2	2		2.8
Trerf1	S421824	9	9092501	9136587	30	3	1.02	1.28	21	40		
Clip1	S420886	12	34049773	34155526	31	2	0.65	1.29	51	99		2.8 (-4.85)
Diablo	S420886	12	34196347	34209619	31	2	0.65	1.29	35	52		2.8 (-4.85)
Tbx3	S694643	12	38155494	38166670	31	10	1.16	1.41	12	25		2.8 (-4.85)
Ptgs2	NA	13	64427288	64433974	32	4	0.18	0.21	NA	366		
Sell	S425490	13	79827342	79845296	33	3	1.18	1.36	NA	146		
Selp	S425490	13	79896091	79922180	33	4	1.18	1.36	NA	110		
F5	S425490	13	79923988	79997285	33	10	1.18	1.36	26	52		
Slc19a2	S425490	13	80017780	80031776	33	8	1.18	1.36	7	10		
Qdpr	S422251	14	70741985	70755600	34	4	1.05	1.17	37	64		
Fhit	S426143: S694236	15	16518688	17329184	35	9	1.07	1.37	94	161		

Tpmt	NA	17	23695915	23714386	36	8	0.14	0.12	14	20		
Hbegf	S420920: S693837	18	29143567	29153944	37	1	0.99	1.35	NA	151	LD6	
Sra1	S420920: S693837	18	29306556	29309783	37	3	0.99	1.35	27	50	LD6	
Gnal	NA	18	63595606	63735803	38	1	0.03	0.02	257	634		
Hsbp1	S422804	19	49589565	49594599	39	6	1.77	2.20	NA	7		
Necab2	S422804	19	49629595	49718718	39	6	1.77	2.20	NA	1		
Mlycd	S422804	19	49637190	49653015	39	12	1.77	2.20	16	34		
Osgin1	S422804	19	49682142	49690478	39	12	1.77	2.20	4	5		
Mbtps1	S422804	19	49751424	49802700	39	6	1.77	2.20	85	140		
Hsdl1	S422804	19	49807249	49814030	39	6	1.77	2.20	1	1		
Lrrc50	S422804	19	49821884	49842233	39	6	1.77	2.20	2	6		
Taf1c	S422804	19	49842372	49848891	39	6	1.77	2.20	37	57		
RT1-S3	NA	20	2815988	2818484	40	4	0.01	0.17	NA	276		(-3.21)
Tff3	S421700	20	9469888	9474605	41	2	1.18	1.37	26	35		
Tff2	S421700	20	9492352	9510558	41	5	1.18	1.37	15	26		
Tff1	S421700	20	9526674	9530534	41	8	1.18	1.37	35	54		

CandidateSNPs: the candidate SNPs selected based on traditional GWAS matched to gene if within 2 Mb distance.
Chr and GeneStart, GeneEnd (Mb): genes' position on chromosome.
Region: candidate genes form clusters along the genome.
SupportScore: assess how well a gene is supported by multiple ranking files from different network and settings.
NodeDegree: the number of other genes connected to each gene in network (Network1 and Network2).
GeneScore(Log_{10}BF): gene score based on Bayes Factor (Log_{10}BF1 or Log_{10}BF2).
LDblocks: the Linakge Disequilibrium detected in Haploview
iHS(XP-EHH): the integrated extended haplotype homozygosity (EHH) in NW population and (the cross population EHH between NW and LH population).

Appendix 5 – Functional annoation analyses of candidate genes from SNP array I (Chapter 3).

Annotation Cluster 1	**Enrichment Score: 2.77**				
Term	Count	PValue	Genes	Fold	Bonferroni
IPR017994:P-type trefoil, chordata	3	0.000	*TFF2, TFF3, TFF1*	97.97	0.05
IPR017957:P-type trefoil, conserved site	3	0.001	*TFF2, TFF3, TFF1*	73.47	0.09
IPR000519:P-type trefoil	3	0.001	*TFF2, TFF3, TFF1*	73.47	0.09
SM00018:PD	3	0.001	*TFF2, TFF3, TFF1*	63.51	0.03
Annotation Cluster 2	**Enrichment Score: 2.32**				
carboxyglutamic acid	3	0.001	*BGLAP, F2, MGP*	72.57	0.08
IPR000294:Gamma-carboxyglutamic acid-rich (GLA) domain	3	0.002	*BGLAP, F2, MGP*	41.99	0.28
gamma-carboxyglutamic acid	3	0.002	*BGLAP, F2, MGP*	41.47	0.24
domain:Gla	3	0.002	*BGLAP, F2, MGP*	40.36	0.30
SM00069:GLA	3	0.003	*BGLAP, F2, MGP*	36.29	0.09
calcium binding	3	0.034	*BGLAP, F2, MGP*	10.01	0.99
Annotation Cluster 3	**Enrichment Score: 2.12**				
GO:0009611~response to wounding	11	0.000	*SELP, ORM1, PTGS2, F5, F3, F2, TGFB3, TFF3, HBEGF, TFF1, BMPR1B*	4.12	0.19
GO:0006954~inflammatory response	6	0.008	*SELP, ORM1, PTGS2, F3, F2, BMPR1B*	4.68	1.00
GO:0006952~defense response	7	0.017	*SELP, ORM1, PTGS2, F3, F2, RSAD2, BMPR1B*	3.33	1.00
Annotation Cluster 4	**Enrichment Score: 2.11**				
GO:0001503~ossification	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	8.77	0.37
GO:0070167~regulation of biomineral formation	4	0.001	*BGLAP, TGFB3, MGP, BMPR1B*	21.84	0.47
GO:0030500~regulation of bone mineralization	4	0.001	*BGLAP, TGFB3, MGP, BMPR1B*	21.84	0.47
GO:0060348~bone development	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	7.78	0.56
GO:0030278~regulation of ossification	5	0.001	*BGLAP, TGFB3, MGP, RSAD2, BMPR1B*	10.80	0.61
GO:0001501~skeletal system development	8	0.001	*FGFR2, BGLAP, PTGS2, TBX3, TGFB3, MGP, RSAD2, BMPR1B*	4.58	0.74
GO:0031214~biomineral formation	3	0.015	*FGFR2, BGLAP, PTGS2*	15.56	1.00
Annotation Cluster 5	**Enrichment Score: 1.69**				
domain:EGF-like	5	0.000	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	29.43	0.00
IPR006209:EGF	5	0.002	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	8.75	0.31
IPR000742:EGF-like, type 3	5	0.008	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	6.12	0.73
IPR006210:EGF-like	5	0.011	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	5.70	0.82
SM00181:EGF	5	0.016	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	4.92	0.42
IPR013032:EGF-like region, conserved site	5	0.025	*SELP, TMEFF1, PTGS2, SELL, HBEGF*	4.37	0.98
Annotation Cluster 6	**Enrichment Score: 1.59**				
GO:0030193~regulation of blood coagulation	3	0.014	*SELP, F2, VKORC1*	16.38	1.00
GO:0050818~regulation of coagulation	3	0.015	*SELP, F2, VKORC1*	15.56	1.00
Annotation Cluster 7	**Enrichment Score: 1.40**				
GO:0042060~wound healing	7	0.001	*F5, F3, F2, TGFB3, TFF3, HBEGF, TFF1*	5.67	0.67

Annotation Cluster 8	**Enrichment Score: 1.33**				
GO:0044421~extracellular region part	12	0.006	*SELP, ORM1, BGLAP, F5, RGD1566296, BCHE, F3, TGFB3, MGP, TFF3, HBEGF, TFF1*	2.50	0.66
Annotation Cluster 9	**Enrichment Score: 1.29**				
GO:0009611~response to wounding	11	0.000	*SELP, ORM1, PTGS2, F5, F3, F2, TGFB3, TFF3, HBEGF, TFF1, BMPR1B*	4.12	0.19
GO:0042060~wound healing	7	0.001	*F5, F3, F2, TGFB3, TFF3, HBEGF, TFF1*	5.67	0.67
disulfide bond	18	0.002	*GGCX, SELP, BGLAP, TMEFF1, PTGS2, SELL, TGFB3, MGP, ORM1, XYLT1, F3, F2, VKORC1, TFF2, TFF3, HBEGF, TFF1, UNC5C*	2.11	0.27
disulfide bond	17	0.003	*GGCX, SELP, BGLAP, TMEFF1, PTGS2, SELL, TGFB3, MGP, ORM1, F3, F2, VKORC1, TFF2, TFF3, HBEGF, TFF1, UNC5C*	2.07	0.43
signal	18	0.006	*SELP, BGLAP, TMEFF1, PTGS2, SELL, PTPRN2, TGFB3, MGP, RT1-S3, ORM1, F3, F2, TFF2, TFF3, HBEGF, TFF1, UNC5C, MBTPS1*	1.93	0.55
GO:0044421~extracellular region part	12	0.006	*SELP, ORM1, BGLAP, F5, RGD1566296, BCHE, F3, TGFB3, MGP, TFF3, HBEGF, TFF1*	2.50	0.66
GO:0005615~extracellular space	10	0.006	*SELP, ORM1, BGLAP, F5, BCHE, F3, TGFB3, MGP, HBEGF, TFF1*	2.87	0.66
signal peptide	18	0.007	*SELP, BGLAP, TMEFF1, PTGS2, SELL, PTPRN2, TGFB3, MGP, RT1-S3, ORM1, F3, F2, TFF2, TFF3, HBEGF, TFF1, UNC5C, MBTPS1*	1.88	0.67
GO:0005576~extracellular region	15	0.015	*SELP, BGLAP, TGFB3, MGP, CALU, ORM1, RGD1566296, F5, BCHE, F3, F2, TFF2, TFF3, HBEGF, TFF1*	1.95	0.92
Annotation Cluster 10	**Enrichment Score: 1.28**				
GO:0001501~skeletal system development	8	0.001	*FGFR2, BGLAP, PTGS2, TBX3, TGFB3, MGP, RSAD2, BMPR1B*	4.58	0.74
Annotation Cluster 11	**Enrichment Score: 1.05**				
Term	Count	PValue	Genes	Fold	Bonferroni
GO:0001503~ossification	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	8.77	0.37
GO:0060348~bone development	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	7.78	0.56
GO:0010033~response to organic substance	11	0.049	*SELP, GNAL, PTGS2, SELL, XYLT1, BCHE, VKORC1, TGFB3, MGP, TFF3, TFF1*	1.93	1.00
Annotation Cluster 12	**Enrichment Score: 1.01**				
GO:0001501~skeletal system development	8	0.001	*FGFR2, BGLAP, PTGS2, TBX3, TGFB3, MGP, RSAD2, BMPR1B*	4.58	0.74
Annotation Cluster 13	**Enrichment Score: 0.98**				
GO:0001503~ossification	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	8.77	0.37
GO:0060348~bone development	6	0.001	*FGFR2, BGLAP, PTGS2, TGFB3, MGP, RSAD2*	7.78	0.56
GO:0001501~skeletal system development	8	0.001	*FGFR2, BGLAP, PTGS2, TBX3, TGFB3, MGP, RSAD2, BMPR1B*	4.58	0.74
GO:0051240~positive regulation of multicellular organismal process	6	0.017	*FGFR2, PTGS2, F2, VKORC1, TGFB3, BMPR1B*	3.94	1.00
Annotation Cluster 14	**Enrichment Score: 0.96**				
Term	Count	PValue	Genes	Fold	Bonferroni
GO:0008083~growth factor activity	5	0.009	*TFF2, TGFB3, HBEGF, OSGIN1, TFF1*	5.96	0.82
Annotation Cluster 15	**Enrichment Score: 0.81**				
GO:0048661~positive regulation of smooth muscle cell proliferation	3	0.041	*FGFR2, PTGS2, HBEGF*	9.15	1.00
Annotation Cluster 16	**Enrichment Score: 0.78**				
GO:0031988~membrane-bounded vesicle	9	0.022	*SELP, BGLAP, CLTA, F5, PTPRN2, AP2S1, TGFB3, TFF3, CALU*	2.52	0.98
GO:0044433~cytoplasmic vesicle part	5	0.025	*SELP, CLTA, F5, AP2S1, TGFB3*	4.37	0.99
GO:0030141~secretory granule	5	0.046	*SELP, F5, PTPRN2, TGFB3, TFF3*	3.63	1.00

GO:0016023~cytoplasmic membrane-bounded vesicle	8	0.049	*SELP, CLTA, F5, PTPRN2, AP2S1, TGFB3, TFF3, CALU*	2.34	1.00
Annotation Cluster 18	**Enrichment Score: 0.77**				
Term	Count	PValue	Genes	Fold	Bonferroni
endoplasmic reticulum	7	0.039	*GGCX, PTGS2, XYLT1, VKORC1, CYP4F1, RSAD2, MBTPS1*	2.72	0.99
GO:0005783~endoplasmic reticulum	10	0.042	*GGCX, PTGS2, XYLT1, BCHE, VKORC1, CYP4F1, MGP, RSAD2, MBTPS1, CALU*	2.10	1.00
Annotation Cluster 19	**Enrichment Score: 0.68**				
GO:0044433~cytoplasmic vesicle part	5	0.025	*SELP, CLTA, F5, AP2S1, TGFB3*	4.37	0.99
GO:0016023~cytoplasmic membrane-bounded vesicle	8	0.049	*SELP, CLTA, F5, PTPRN2, AP2S1, TGFB3, TFF3, CALU*	2.34	1.00
GO:0031410~cytoplasmic vesicle	8	0.092	*SELP, CLTA, F5, PTPRN2, AP2S1, TGFB3, TFF3, CALU*	2.02	1.00
Annotation Cluster 20	**Enrichment Score: 0.64**				
Term	Count	PValue	Genes	Fold	Bonferroni
GO:0005788~endoplasmic reticulum lumen	3	0.038	*PTGS2, BCHE, CALU*	9.45	1.00

Appendix 6 – Candidate genes with expression changes clustered in 21 regions along the genome (Chapter 4).

Targets	Clustered Regions	Chr	GenePos	ProbeID	GGIRankScore	CTDRankScore
Cyp2t1	1	1	82228606	1368265_at	NA	0.551
Fbl	1	1	83271967	1388528_at	0.757	0.578
Supt5h	1	1	83404787	1372094_at	NA	0.623
Zfp36	1	1	83486643	1387870_at	0.750	0.563
Eif3k	1	1	84083646	1372735_at	0.752	0.329
Tyrobp	1	1	85365371	1374730_at	0.870	0.655
Tbcb	1	1	85490553	1371885_at	0.780	NA
Rbm42	1	1	85692754	1371502_at	0.718	0.555
Fxyd1	1	1	86095551	1369960_at	0.717	0.467
Gpi	1	1	86658839	1371392_at	0.696	0.478
Tufm	2	1	185631333	1371962_at	NA	0.394
Coro1a	2	1	185852742	1369964_at	0.850	0.671
Aldoa	2	1	185970658	1367617_at	0.769	0.533
Cdipt	2	1	186153278	1387900_at	0.843	0.597
Znf553	2	1	186498827	1383260_at	NA	NA
Znf771	2	1	186507474	1383260_at	NA	0.531
Dctpp1	2	1	186519957	1370308_at	NA	0.528
Pycard	2	1	187276089	1389873_at	0.802	0.625
Psmd13	3	1	201044403	1371617_at	0.735	0.485
Rplp2	3	1	201635581	1371340_at	NA	0.377
Ap2a2	3	1	201741585	1371493_at	0.616	0.435
Ctsd	3	1	202619669	1367651_at	0.830	0.586
Nadsyn1	3	1	204187196	1379472_at	0.619	0.438
Cpt1a	3	1	205852746	1386946_at	0.655	10.770
Cdk2ap2	3	1	206681430	1373901_at	NA	0.564

Pitpnm1	3	1	206684110	1389347_at	0.678	0.530
Coro1b	3	1	206733359	1398769_at	0.812	0.637
Ctsf	3	1	207469235	1373258_at	NA	0.274
Brms1	3	1	207661537	1373428_at	0.728	0.491
Cfl1	3	1	208133876	1370184_at	0.759	0.610
Fau	3	1	208695065	1371316_at	NA	0.460
Ehd1	3	1	209055078	1388623_at	0.676	0.501
Sf1	3	1	209146177	1398339_at	0.746	0.498
Prdx5	3	1	209583414	1367677_at	0.856	0.661
Ppp1r14b	3	1	209648657	1389815_at	NA	0.649
Ubxn1	3	1	211540964	1373299_at	0.660	NA
Mta2	3	1	211618426	1373103_at	0.685	0.521
Tmem216	3	1	213037899	1382049_at	NA	0.457
Ddb1	3	1	213095276	1399162_a_at	0.812	0.492
Prpf19	3	1	213390564	1398867_at	NA	0.460
Slc25a28	4	1	247159998	1392978_at	0.704	NA
Bloc1s2	4	1	249086242	1398473_at	0.769	0.593
Fbxw4	4	1	250746869	1373211_at	0.687	0.333
Cuedc2	4	1	251546667	1376368_at	NA	NA
Actr1a	4	1	251592429	1371741_at	0.679	0.397
Taf5	4	1	252391049	1376050_at	0.724	0.563
Rps3a	5	2	178092354	1367606_at	0.664	0.351
Hdgf	5	2	179991032	1367817_at	0.810	0.523
Scamp3	5	2	181249544	1370921_at	NA	0.501
Flad1	5	2	181598343	1373758_at	0.703	NA
Il6ra	5	2	182078051	1386987_at	NA	0.607
Snapap	5	2	182762319	1376366_at	0.730	NA
Fitm2	6	3	154375988	1394392_at	NA	0.490
Sdc4	6	3	155446138	1367721_at	0.723	0.616
LOC685079	6	3	155482350	1367543_at	0.763	NA
Pltp	6	3	155872122	1391435_at	0.707	0.545
Slc13a3	6	3	156447894	1368047_at	0.641	0.320
Ddx27	6	3	158165670	1385471_at	0.651	0.364
Mast2	7	5	136573518	1388399_at	0.753	0.582
LOC500532	7	5	137578017	1371721_at	0.851	NA
Atp6v0b	7	5	138349548	1398930_at	NA	0.426
Ipo13	7	5	138359936	1367876_at	0.766	0.583
Ebna1bp2	7	5	139095875	1388397_at	0.674	0.477
Ppcs	7	5	140008436	1388756_at	0.698	0.427
Zfp347	8	7	8938240	1368726_a_at	0.810	0.546
Dohh	8	7	9806009	1372677_at	0.774	0.403
Sgta	8	7	10137703	1388467_at	0.627	0.374
Slc39a3	8	7	10160746	1377933_at	0.714	0.535
Oaz1	8	7	10394495	1372241_at	0.759	0.461
Ap3d1	8	7	10480810	1367549_a_at	0.731	0.357
Mbd3	8	7	10824397	1371345_at	NA	0.617
Atp5d	8	7	11072820	1370278_at	0.655	0.486
Sbno2	8	7	11117835	1392107_at	0.830	0.440

Wdr18	8	7	11253443	1375439_at	0.623	0.453
Ppap2c	8	7	11694272	1388913_at	0.778	0.622
Brd4	8	7	12772428	1372351_at	0.704	0.339
Cyp4f6	8	7	13609103	1387916_at	0.627	0.431
Cdk4	9	7	67016944	1369950_at	NA	0.652
Tspan31	9	7	67019197	1372739_at	0.612	0.463
Sdc2	9	7	68290768	1382189_at	0.330	0.243
Rpl30	9	7	69886763	1398774_at	NA	0.517
Hrsp12	9	7	69930466	1368060_at	0.720	0.429
Ly6e	10	7	113151828	1388347_at	0.804	0.630
Eef1d	10	7	113870559	1388134_at	0.731	0.499
Puf60	10	7	114101195	1367464_at	0.782	NA
LOC100362108	10	7	114264088	1376056_at	0.751	NA
Oplah	10	7	114326739	1368091_at	0.721	0.491
Exosc4	10	7	114374913	1371670_at	0.643	0.514
Maf1	10	7	114402298	1371374_at	0.767	0.607
Bop1	10	7	114500159	1392910_at	0.777	0.547
Dgat1	10	7	114552056	1367915_at	0.735	0.579
MGC94207	10	7	114766704	1372562_at	NA	NA
Eif3d	10	7	115877161	1388568_at	0.836	0.524
Cdc42ep1	10	7	116807762	1389157_at	0.836	0.641
Tuba1b	11	7	137706856	1367579_a_at	0.801	0.619
Tuba1a	11	7	137729408	1367579_a_at	0.805	0.640
Tuba1c	11	7	137809847	1367579_a_at	0.801	0.626
Gpd1	11	7	138458875	1371363_at	0.402	0.362
RGD1359310	11	7	140029813	1372026_at	0.660	NA
Oaf	12	8	46224862	1388425_at	0.609	0.483
Mizf	12	8	47275524	1391849_at	0.690	NA
Hmbs	12	8	47314235	1386983_at	0.773	0.608
Slc37a4	12	8	47363896	1386960_at	0.668	0.585
Fxyd2	12	8	48379075	1387799_at	NA	0.499
Sidt2	12	8	48909652	1388828_at	0.707	0.546
Rassf1	13	8	112798863	1373989_at	0.699	0.334
Ifrd2	13	8	112834734	1373068_at	0.694	0.464
Slc38a3	13	8	112898379	1370824_at	0.630	0.499
Map4	13	8	114354729	1373268_at	0.679	0.431
Dhx30	13	8	114438390	1388595_at	0.857	0.587
Lrrfip2	13	8	115509800	1379282_at	0.780	0.405
Tmem132e	14	10	70584564	1384771_at	NA	NA
Slfn2	14	10	71236408	1377916_at	NA	0.614
Ap2b1	14	10	71386233	1367704_at	0.727	0.549
Ccl6	14	10	71659632	1389123_at	0.538	0.735
Usp32	14	10	73256761	1373962_at	0.347	0.271
Galk1	15	10	106116775	1389248_at	0.747	0.523
H3f3b	15	10	106130318	1398888_at	NA	0.527
Prpsap1	15	10	106582382	1367750_at	0.827	0.709
Rbm25	15	10	107711516	1392936_at	0.693	0.378
Usp36	15	10	108264874	1374940_at	NA	0.300

Lgals3bp	15	10	108388185	1387946_at	0.682	0.549
Nploc4	15	10	109809121	1393244_at	0.753	0.518
Hgs	15	10	109868142	1367840_at	0.667	0.359
Thoc4	15	10	109984500	1374897_at	0.742	0.552
Sectm1b	15	10	110271329	1376976_at	0.783	0.510
Chrd	16	11	82401720	1380285_at	0.650	0.515
Eif4g1	16	11	82452006	1388322_at	NA	0.440
Psmd2	16	11	82478179	1398858_at	0.831	0.496
Abcf3	16	11	82570496	1371456_at	0.732	0.556
Ap2m1	16	11	82585474	1398765_at	0.839	0.533
Klhl24	16	11	83097219	1383110_at	NA	0.425
Setd8	17	12	33253847	1375928_at	NA	0.464
Clip1	17	12	34049773	1367993_at	0.632	0.489
Orai1	17	12	34631272	1390031_at	0.733	0.567
Pptc7	17	12	35445316	1388522_at	NA	0.377
Plbd2	17	12	37260666	1375173_at	0.734	0.540
Tcn2	18	14	84579191	1367765_at	0.747	0.587
Ewsr1	18	14	85730484	1389289_at	0.690	0.544
Dbnl	18	14	86450408	1368274_at	0.701	0.543
Ykt6	18	14	86619828	1372505_at	0.545	0.471
Ddx56	18	14	86985895	1389389_at	0.769	0.508
Purb	18	14	87231825	1374205_at	0.717	0.369
Ccm2	18	14	87329085	1375282_at	0.642	0.490
Mbl1	19	16	17591820	1387765_at	0.724	0.505
LOC688495	19	16	17685244	1399080_at	NA	NA
Cherp	19	16	17785631	1372705_at	NA	0.428
Fam125a	19	16	18726945	1388457_at	0.709	0.551
Pgls	19	16	18791241	1374523_at	0.692	0.532
LOC498606	19	16	19413529	1389156_at	0.687	NA
Sf4	19	16	19835922	1376164_at	0.695	0.398
Atp13a1	19	16	20100482	1371925_at	0.735	0.580
Mlycd	**20**	19	49637190	1367638_at	0.690	0.412
Cotl1	**20**	19	50096633	1388596_at	0.810	0.617
Fbxo31	**20**	19	51823606	1372600_at	NA	0.470
Zc3h18	**20**	19	52654167	1391455_at	NA	0.533
Rpl13	**20**	19	53437633	1386858_at	0.697	0.583
Tcf25	**20**	19	53700976	1391346_at	0.522	0.292
Znrd1	**21**	20	1687850	1388590_at	0.705	0.548
Rnf39	**21**	20	1697530	1368662_at	0.618	0.490
Abcf1	**21**	20	2953601	1398876_at	0.766	0.569
RGD1309543	**21**	20	3025507	1373262_at	0.824	NA
RT1-CE5	**21**	20	3509597	1370972_x_at	0.545	0.413

Appendix 7 – Functional annotation analyses of candidate genes from microarray data (Chapter 4).

Annotation Cluster 1	**Enrichment Score: 2.89**				
Term	Count	PValue	Genes	Fold	Bonferroni
GO:0043232~intracellular non-membrane-bounded organelle	33	0.000	*ABCF1, TUFM, ALDOA, RPL13, RPLP2, ZNRD1, BOP1, FAM125A, RPL30, CDC42EP1, BLOC1S2, RPS3A, ACTR1A, FAU, DHX30, TUBA1A, TUBA1B, TUBA1C, DBNL, TAF5, HMBS, MBD3, COTL1, FBL, PURB, CORO1B, CORO1A, DDX56, MAST2, CFL1, CLIP1, MAP4, H3F3B, SLC13A3*	1.86	0.08
GO:0043228~non-membrane-bounded organelle	33	0.000	*ABCF1, TUFM, ALDOA, RPL13, RPLP2, ZNRD1, BOP1, FAM125A, RPL30, CDC42EP1, BLOC1S2, RPS3A, ACTR1A, FAU, DHX30, TUBA1A, TUBA1B, TUBA1C, DBNL, TAF5, HMBS, MBD3, COTL1, FBL, PURB, CORO1B, CORO1A, DDX56, MAST2, CFL1, CLIP1, MAP4, H3F3B, SLC13A3*	1.86	0.08
GO:0005856~cytoskeleton	17	0.018	*ALDOA, DBNL, TAF5, COTL1, FAM125A, CORO1B, CORO1A, MAST2, CDC42EP1, BLOC1S2, ACTR1A, CFL1, MAP4, CLIP1, SLC13A3, TUBA1A, TUBA1B, TUBA1C*	1.86	0.98
Annotation Cluster 2	**Enrichment Score: 1.80**				
domain:ADF-H	3	0.004	*DBNL, CFL1, COTL1*	31.05	0.66
SM00102:ADF	3	0.005	*DBNL, CFL1, COTL1*	27.33	0.28
IPR002108:Actin-binding, cofilin/tropomyosin type	3	0.005	*DBNL, CFL1, COTL1*	27.09	0.78
actin-binding	5	0.059	*CORO1B, DBNL, CORO1A, CFL1, COTL1*	3.37	1.00
Annotation Cluster 3	**Enrichment Score: 1.39**				
GO:0006414~translational elongation	7	0.000	*TUFM, RPL30, RPS3A, RPL13, RPLP2, FAU, EEF1D*	8.03	0.22
protein biosynthesis	7	0.008	*TUFM, EIF3D, RPL30, RPL13, RPLP2, FAU, EEF1D*	4.00	0.81
rno03010:Ribosome	5	0.012	*RPL30, RPS3A, RPL13, RPLP2, FAU*	5.52	0.59
GO:0044445~cytosolic part	5	0.021	*BLOC1S2, RPS3A, RPL13, PRDX5, FAU*	4.71	0.99
GO:0030529~ribonucleoprotein complex	11	0.025	*ABCF1, SF4, PRPF19, RPL30, RPS3A, RPL13, SF1, RPLP2, FAU, BOP1, FBL*	2.22	1.00
ribonucleoprotein	6	0.037	*RPL30, RPS3A, RPL13, RPLP2, FAU, FBL*	3.22	1.00
GO:0006412~translation	9	0.041	*ABCF1, TUFM, EIF3D, RPL30, RPS3A, RPL13, RPLP2, FAU, EEF1D*	2.29	1.00
Annotation Cluster 4	**Enrichment Score: 1.30**				
GO:0050690~regulation of defense response to virus by virus	3	0.002	*AP2A2, AP2B1, AP2M1*	40.85	0.91
GO:0030139~endocytic vesicle	5	0.003	*AP2A2, AP2B1, CORO1A, EHD1, AP2M1*	8.41	0.48
IPR002553:Clathrin/coatomer adaptor, adaptin-like, N-terminal	3	0.003	*AP2A2, AP2B1, AP3D1*	34.83	0.59
GO:0030666~endocytic vesicle membrane	4	0.004	*AP2A2, AP2B1, CORO1A, AP2M1*	12.07	0.62
GO:0009898~internal side of plasma membrane	7	0.005	*AP2A2, AP2B1, HGS, SNAPAP, EHD1, YKT6, AP2M1*	4.43	0.67
IPR011989:Armadillo-like helical	5	0.006	*AP2A2, AP2B1, DOHH, IPO13, AP3D1*	6.66	0.86
GO:0050688~regulation of defense response to virus	3	0.008	*AP2A2, AP2B1, AP2M1*	21.79	1.00
GO:0030117~membrane coat	4	0.012	*AP2A2, AP2B1, AP3D1, AP2M1*	8.21	0.94
GO:0048475~coated membrane	4	0.012	*AP2A2, AP2B1, AP3D1, AP2M1*	8.21	0.94
GO:0030119~AP-type membrane coat adaptor complex	3	0.017	*AP2A2, AP2B1, AP2M1*	14.66	0.98
GO:0030131~clathrin adaptor complex	3	0.017	*AP2A2, AP2B1, AP2M1*	14.66	0.98

GO:0043900~regulation of multi-organism process	3	0.020	*AP2A2, AP2B1, AP2M1*	13.62	1.00
rno04144:Endocytosis	7	0.020	*FAM125A, AP2A2, AP2B1, RT1-CE5, HGS, EHD1, AP2M1*	3.19	0.78
GO:0002831~regulation of response to biotic stimulus	3	0.023	*AP2A2, AP2B1, AP2M1*	12.57	1.00
GO:0008565~protein transporter activity	4	0.025	*AP2A2, AP2B1, IPO13, AP3D1*	6.35	1.00
protein transport	8	0.028	*PITPNM1, FAM125A, AP2A2, AP2B1, IPO13, HGS, YKT6, AP2M1*	2.70	1.00
GO:0015031~protein transport	11	0.028	*PITPNM1, FAM125A, AP2A2, SCAMP3, AP2B1, IPO13, CFL1, HGS, AP3D1, YKT6, AP2M1*	2.18	1.00
coated pit	3	0.029	*AP2A2, AP2B1, AP2M1*	11.13	1.00
GO:0045184~establishment of protein localization	11	0.029	*PITPNM1, FAM125A, AP2A2, SCAMP3, AP2B1, IPO13, CFL1, HGS, AP3D1, YKT6, AP2M1*	2.16	1.00
GO:0005905~coated pit	3	0.034	*AP2A2, AP2B1, AP2M1*	10.26	1.00
GO:0046907~intracellular transport	10	0.035	*AP2A2, AP2B1, IPO13, CFL1, HGS, AP3D1, THOC4, EHD1, YKT6, AP2M1*	2.21	1.00
GO:0030118~clathrin coat	3	0.036	*AP2A2, AP2B1, AP2M1*	9.93	1.00
GO:0016192~vesicle-mediated transport	9	0.046	*DBNL, AP2A2, AP2B1, CORO1A, AP3D1, SNAPAP, EHD1, YKT6, AP2M1*	2.24	1.00
GO:0006886~intracellular protein transport	7	0.046	*AP2A2, AP2B1, IPO13, CFL1, HGS, AP3D1, AP2M1*	2.67	1.00
Annotation Cluster 5	**Enrichment Score: 1.23**				
GO:0006733~oxidoreduction coenzyme metabolic process	4	0.012	*GPD1, PGLS, NADSYN1, FLAD1*	8.38	1.00
GO:0051186~cofactor metabolic process	6	0.037	*GPD1, PGLS, MLYCD, HMBS, NADSYN1, FLAD1*	3.24	1.00
Annotation Cluster 6	**Enrichment Score: 1.23**				
GO:0048029~monosaccharide binding	5	0.002	*ALDOA, MBL1, GPI, GALK1, PGLS*	9.92	0.37
GO:0019318~hexose metabolic process	7	0.011	*ALDOA, GPI, GPD1, GALK1, PGLS, SLC37A4, CPT1A*	3.76	1.00
GO:0005996~monosaccharide metabolic process	7	0.018	*ALDOA, GPI, GPD1, GALK1, PGLS, SLC37A4, CPT1A*	3.32	1.00
GO:0006006~glucose metabolic process	6	0.022	*ALDOA, GPI, GPD1, PGLS, SLC37A4, CPT1A*	3.71	1.00
rno00030:Pentose phosphate pathway	3	0.027	*ALDOA, GPI, PGLS*	11.45	0.87
GO:0044275~cellular carbohydrate catabolic process	4	0.028	*ALDOA, GPI, GPD1, PGLS*	6.05	1.00
GO:0046164~alcohol catabolic process	4	0.028	*ALDOA, GPI, GPD1, PGLS*	6.05	1.00
GO:0030246~carbohydrate binding	8	0.044	*ALDOA, MBL1, GPI, GALK1, CDIPT, PGLS, HDGF, CHRD*	2.45	1.00
Annotation Cluster 7	**Enrichment Score: 1.22**				
repeat:WD 5	5	0.023	*PRPF19, CORO1B, CORO1A, WDR18, BOP1*	4.59	1.00
repeat:WD 4	5	0.026	*PRPF19, CORO1B, CORO1A, WDR18, BOP1*	4.42	1.00
repeat:WD 3	5	0.029	*PRPF19, CORO1B, CORO1A, WDR18, BOP1*	4.26	1.00
repeat:WD 2	5	0.030	*PRPF19, CORO1B, CORO1A, WDR18, BOP1*	4.21	1.00
repeat:WD 1	5	0.030	*PRPF19, CORO1B, CORO1A, WDR18, BOP1*	4.21	1.00
Annotation Cluster 8	**Enrichment Score: 1.14**				
GO:0009165~nucleotide biosynthetic process	7	0.008	*ATP5D, ALDOA, ATP13A1, NADSYN1, FLAD1, PRPSAP1, ATP6V0B*	3.95	1.00
GO:0034404~nucleobase, nucleoside and nucleotide biosynthetic process	7	0.010	*ATP5D, ALDOA, ATP13A1, NADSYN1, FLAD1, PRPSAP1, ATP6V0B*	3.81	1.00
GO:0034654~nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process	7	0.010	*ATP5D, ALDOA, ATP13A1, NADSYN1, FLAD1, PRPSAP1, ATP6V0B*	3.81	1.00
GO:0044271~nitrogen compound biosynthetic	8	0.030	*ATP5D, ALDOA, ATP13A1, HMBS, NADSYN1, FLAD1, PRPSAP1, ATP6V0B*	2.65	1.00
Annotation Cluster 9	**Enrichment Score: 1.06**				

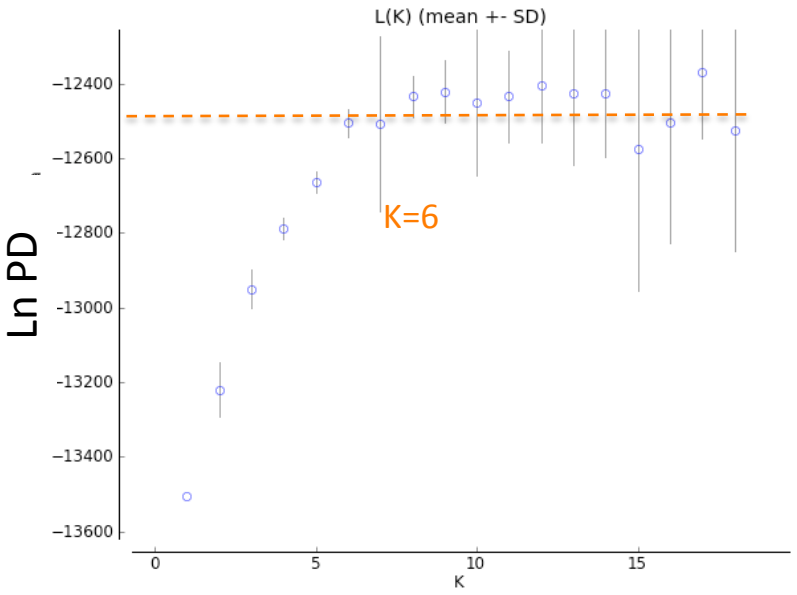
GO:0042802~identical protein binding	12	0.026	*MBL1, GPD1, DBNL, CORO1A, DDX56, BLOC1S2, RASSF1, PYCARD, CLIP1, PRPSAP1, CPT1A, IL6RA*	2.10	1.00
Annotation Cluster 10	**Enrichment Score: 1.05**				
GO:0016887~ATPase activity	8	0.015	*DDX27, ABCF1, ATP5D, FXYD2, ABCF3, DDX56, ATP13A1, DHX30*	3.07	0.99
GO:0042623~ATPase activity, coupled	6	0.044	*DDX27, ATP5D, FXYD2, DDX56, ATP13A1, DHX30*	3.09	1.00
Annotation Cluster 11	**Enrichment Score: 1.04**				
GO:0006511~ubiquitin-dependent protein catabolic process	6	0.018	*PSMD13, DDB1, PSMD2, FBXO31, USP36, USP32*	3.94	1.00
GO:0043632~modification-dependent macromolecule catabolic process	7	0.045	*CUEDC2, PSMD13, DDB1, PSMD2, FBXO31, USP36, USP32*	2.69	1.00
GO:0019941~modification-dependent protein catabolic process	7	0.045	*CUEDC2, PSMD13, DDB1, PSMD2, FBXO31, USP36, USP32*	2.69	1.00
Annotation Cluster 12	**Enrichment Score: 1.03**				
GO:0008092~cytoskeletal protein binding	10	0.013	*CORO1B, DBNL, CORO1A, MAST2, BLOC1S2, CFL1, CLIP1, SDC4, COTL1, SDC2*	2.64	0.98
cytoskeleton	9	0.014	*CORO1B, DBNL, FAM125A, CORO1A, CDC42EP1, BLOC1S2, ACTR1A, CFL1, COTL1*	2.80	0.95
GO:0005856~cytoskeleton	17	0.018	*ALDOA, DBNL, TAF5, COTL1, FAM125A, CORO1B, CORO1A, MAST2, CDC42EP1, BLOC1S2, ACTR1A, CFL1, MAP4, CLIP1, SLC13A3, TUBA1A, TUBA1B, TUBA1C*	1.86	0.98
Annotation Cluster 13	**Enrichment Score: 0.98**				
GO:0006639~acylglycerol metabolic process	4	0.014	*CDIPT, DGAT1, SLC37A4, CPT1A*	7.78	1.00
GO:0006638~neutral lipid metabolic process	4	0.016	*CDIPT, DGAT1, SLC37A4, CPT1A*	7.51	1.00
GO:0006662~glycerol ether metabolic process	4	0.016	*CDIPT, DGAT1, SLC37A4, CPT1A*	7.39	1.00
GO:0018904~organic ether metabolic process	4	0.018	*CDIPT, DGAT1, SLC37A4, CPT1A*	7.14	1.00
GO:0006006~glucose metabolic process	6	0.022	*ALDOA, GPI, GPD1, PGLS, SLC37A4, CPT1A*	3.71	1.00
GO:0005624~membrane fraction	7	0.697	*GPI, DGAT1, CYP4F6, SLC37A4, SLC13A3, SNAPAP, CPT1A*	1.00	1.00
GO:0005626~insoluble fraction	7	0.745	*GPI, DGAT1, CYP4F6, SLC37A4, SLC13A3, SNAPAP, CPT1A*	0.95	1.00
Annotation Cluster 14	**Enrichment Score: 0.92**				
GO:0015630~microtubule cytoskeleton	9	0.028	*FAM125A, MAST2, BLOC1S2, ACTR1A, CLIP1, MAP4, SLC13A3, TUBA1A, TUBA1B, TUBA1C*	2.48	1.00
Annotation Cluster 16	**Enrichment Score: 0.87**				
GO:0016887~ATPase activity	8	0.015	*DDX27, ABCF1, ATP5D, FXYD2, ABCF3, DDX56, ATP13A1, DHX30*	3.07	0.99
GO:0042623~ATPase activity, coupled	6	0.044	*DDX27, ATP5D, FXYD2, DDX56, ATP13A1, DHX30*	3.09	1.00
Annotation Cluster 19	**Enrichment Score: 0.83**				
GO:0006396~RNA processing	8	0.046	*ZFP36, SF4, PRPF19, DDX56, THOC4, BOP1, RBM25, FBL*	2.42	1.00
Annotation Cluster 20	**Enrichment Score: 0.79**				
GO:0006396~RNA processing	8	0.046	*ZFP36, SF4, PRPF19, DDX56, THOC4, BOP1, RBM25, FBL*	2.42	1.00
Annotation Cluster 21	**Enrichment Score: 0.69**				
GO:0006812~cation transport	10	0.035	*ATP5D, ORAI1, FXYD2, SLC38A3, CORO1A, ATP13A1, SLC13A3, TCN2, SLC39A3, ATP6V0B*	2.21	1.00
Annotation Cluster 24	**Enrichment Score: 0.51**				
GO:0016887~ATPase activity	8	0.015	*DDX27, ABCF1, ATP5D, FXYD2, ABCF3, DDX56, ATP13A1, DHX30*	3.07	0.99
Annotation Cluster 30	**Enrichment Score: 0.31**				
GO:0030246~carbohydrate binding	8	0.044	*ALDOA, MBL1, GPI, GALK1, CDIPT, PGLS, HDGF, CHRD*	2.45	1.00

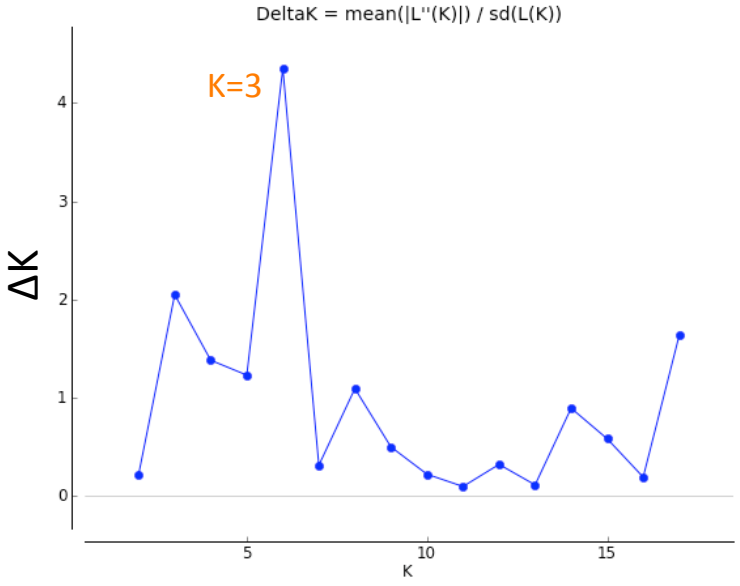
Appendix 8 – Genes involved in warfarin interactive pathways in human.

Genes	Function	Reference
	Genes reported to be connected to *VKORC1* in STRING database	
VKORC1	Encodes a protein of the vitamin K 2,3-epoxide reductase complex (VKOR), explain 30% dose variance	(Li et al. 2004; Rost et al. 2004)
CYP2C9	Dose predictor *CYP2C9* in human, explain ~ 10% dose variance	(Takeuchi et al. 2009)
CYP4F2	Genetic determinant of warfarin dose, explain ~1.5% dose variance	(Takeuchi et al. 2009)
EPHX1	Associated with the maintenance dose of warfarin in elderly populations	(Pautas et al. 2009)
CALU	Encodes calumenin, inhibits both the GGCX and the VKOR; suggested to contribute to warfarin resistance by over-expression in north America	(Wajih et al. 2004)
GGCX	Catalyzes the posttranslational modification of glutamate to gamma-carboxyglutamate (Gla) in vitamin K-dependent proteins	(Markussen et al. 2007b)
NQO1	Encodes a warfarin sensitive NAD(P)H quinone dehydrogenase, involved in vitamin K reduction pathway, potentially reduce dietary vitamin K	(Markussen et al. 2007b)
APOE	Apolipoprotein E, clear Vitamin K1, absorbed along with dietary fat	(Wadelius et al. 2007)
	Vitamin K dependent genes involved in blood coagulation	
F2	Blood coagulation factor II (thrombin)	(Stafford 2005)
F7	Blood coagulation factor VII (serum prothrombin conversion accelerator)	(Stafford 2005)
F9	Blood coagulation factor IX	(Stafford 2005)
F10	Blood coagulation factor X	(Stafford 2005)
PROS1	Vitamin K dependent protein S	(Stafford 2005)
PROC	Vitamin K dependent anticoagulant protein C	(Stafford 2005)
PROZ	Vitamin K dependent protein Z	(Stafford 2005)
SERPRINC1	inhibits the Vitamin K dependent clotting factors	
	Vitamin K dependent genes involved in cell cycle regulation	
GAS6	Encodes a unique vitamin K-dependent autocrine growth factor for mesangial cell. Affect vascular smooth muscle cell movement and apoptosis.	(Yanagita 2004; Danziger 2008)
	Vitamin K dependent genes involved in bone metabolism	
MGP	Encodes vitamin K-dependent protein (matrix Gla), which prevents vascular calcification. Its effect on calcification is determined by the relative amounts of MGP and BMP-2.	(Danziger 2008) (Zebboudj, Shin, and Bostrom 2003)
BGLAP	Vitamin K-dependent protein gamma-carboxyglutamyl (bone Gla)	(Suttie 1993)
AXL	Receptor of *GAS6*	(Yanagita 2004)
STAT3	A downstream signaling molecule of the GAS6/AXL pathway	(Yanagita 2004)
	Other interesting genes	
FGG	Combines with alpha and beta polypeptides to form the mature fibrinogen molecule, which is an important component of the coagulation cascade;	(Ivaskevicius et al. 2005)
CYP1A2, CYP3A4, CYP1A1, CYP3A5 CYP2C8	Metabolize R-warfarin or S-warfarin	(Wadelius et al. 2007).
ORM1, ORM2 ABCB1	Warfarin transport in and out of the liver	(Wadelius et al. 2007).
CACNA1C	Subunit of an L-type calcium channel that may be involved in bone metabolism and compensatory renal growth	(Cooper et al. 2008)

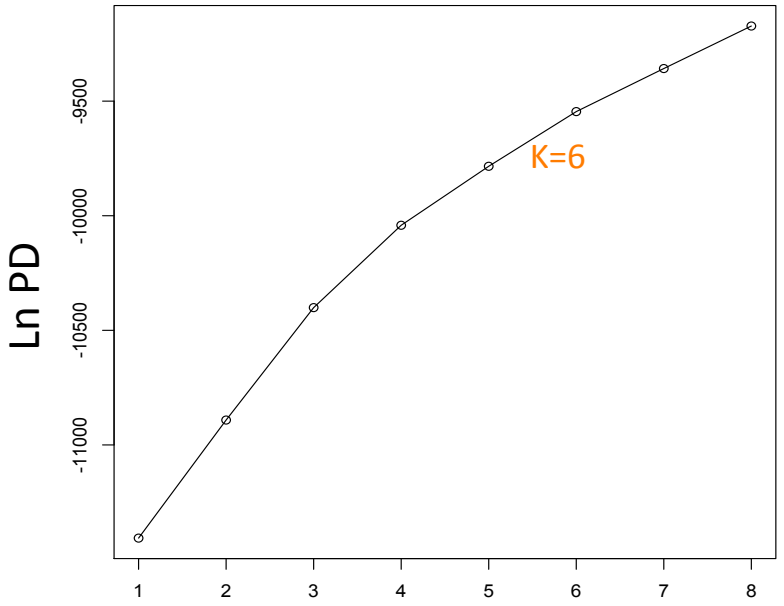
Appendix 9 - Population structure analysis: criteria for choosing optimal cluster number (K)

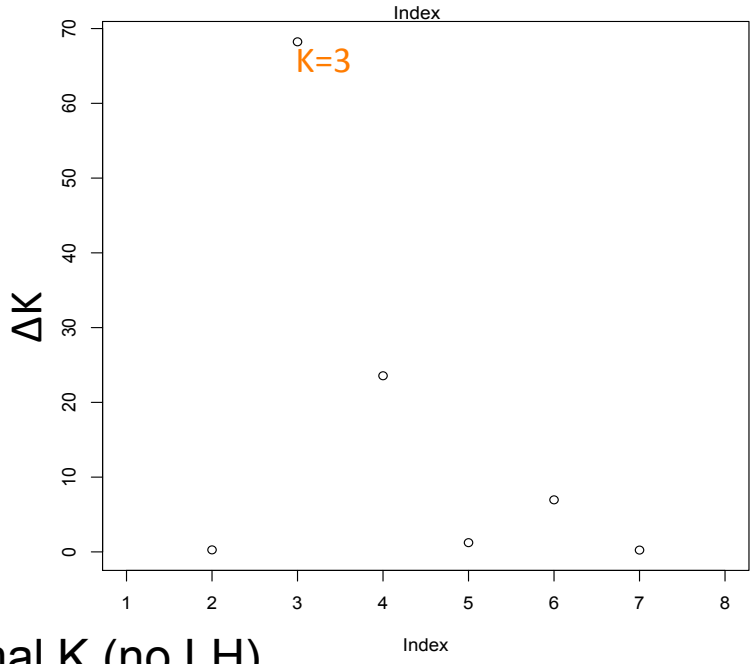
STRUCTURE





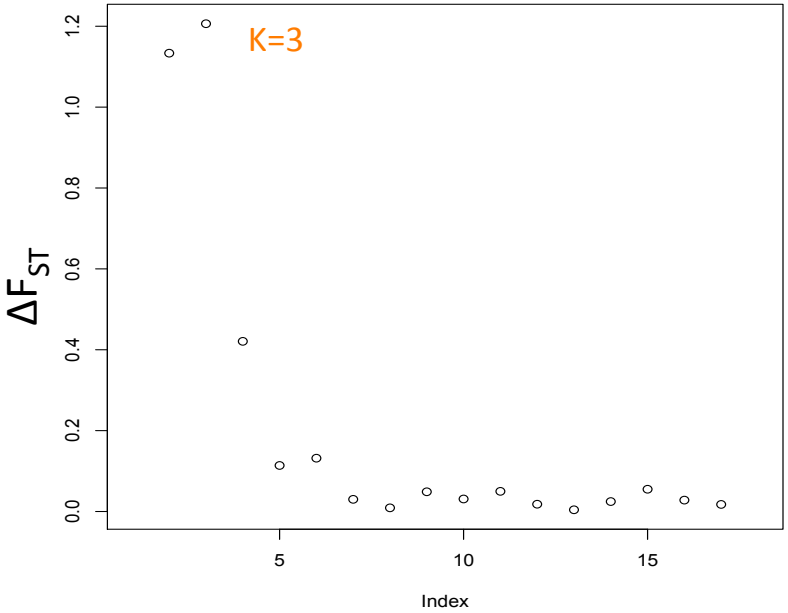
INSTRUCT (inbreeding)

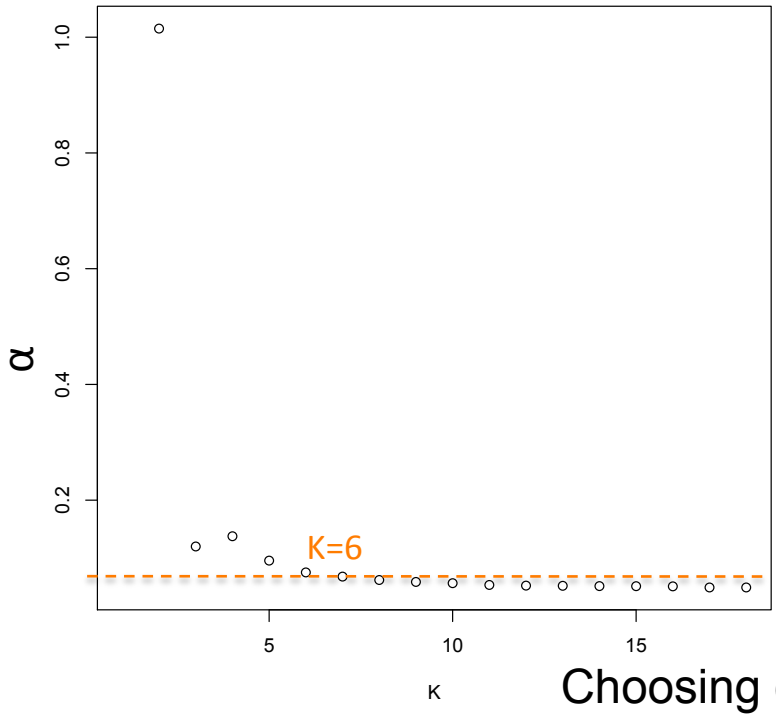




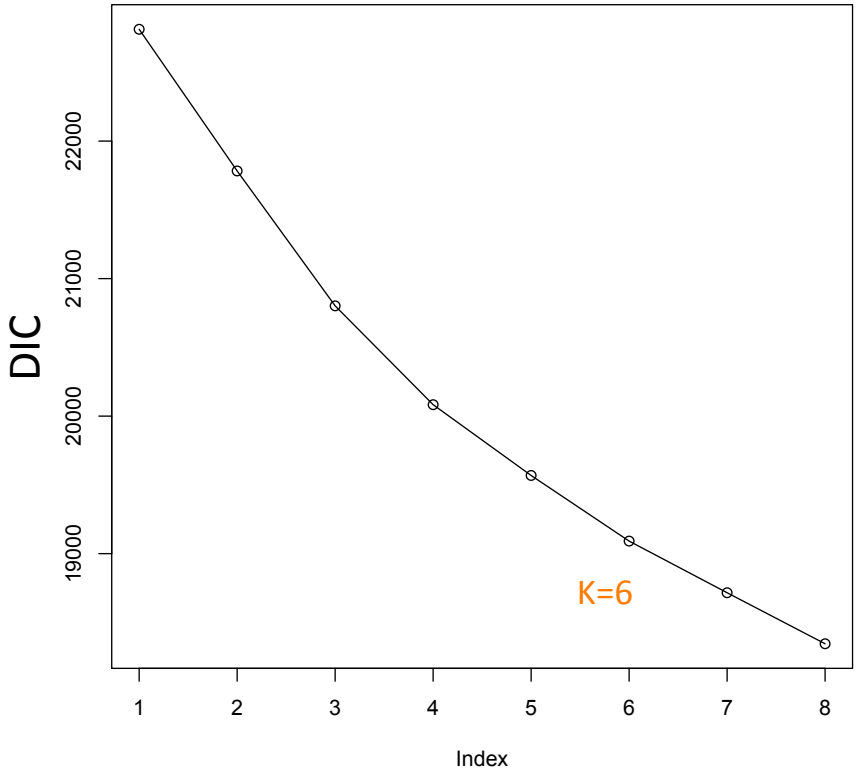
Choosing optimal K (no LH)

STRUCTURE





INSTRUCT (inbreeding)



Choosing optimal K (no LH)

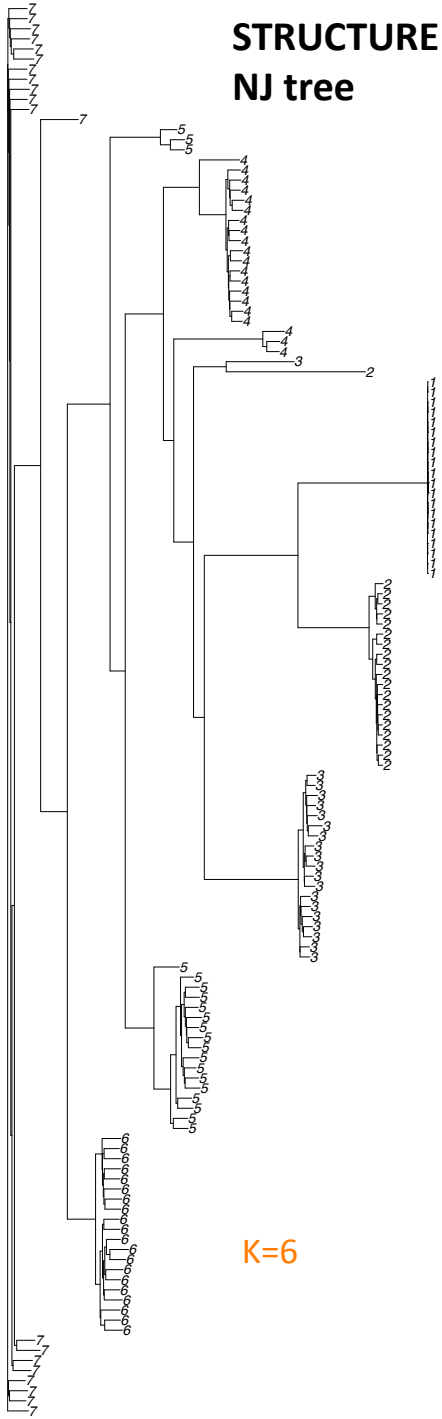
STRUCTURE

K	Max Corr	Sig
2	0.767	N
3	0.970	N
4	0.987	N
5	0.982	N
6	0.998	Y
7	0.997	Y
8	0.991	Y
9	0.990	N
10	0.986	N
11	0.987	N
12	0.985	N
13	0.987	N
14	0.986	N

INSTRUCT (inbreeding)

K	Max Corr	Sig
2	0.646	N
3	0.999	Y
4	0.999	Y
5	0.998	Y
6	0.999	Y
7	0.999	Y
8	0.999	Y

Choosing optimal K (no LH)



Choosing optimal K based on several criteria

Criteria	STRUCTURE	INSTRUCT
LnPD	6	6
ΔK	3	3
ΔF_{ST}	3	3
alpha	6	na
DIC	na	6
AverMaxCorr	6	3
NJtree	6	na